The Synthesis of Novel Antibacterial Proteins in the *Chlamydomonas reinhardtii* Chloroplast

Henry Nicholas Taunt

UCL

A thesis submitted for the degree of

Doctor of Philosophy

September 2013

I, Henry Nicholas Taunt, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....

For Ria,

without whom this

thesis would not have happened

Acknowledgements

Firstly I would like to thank my supervisor, Dr Saul Purton, for teaching me to be a researcher and giving me the freedom to explore my own ideas, irrespective of their removal from the metaphorical box.

I would like to thank all the members of the Purton group with whom I have worked throughout this PhD: in particular Dr Chloe Economou, for helping me find my feet on first arriving in the lab; Laura Stoffels, with whom I worked closely on development of lysin activity assays; Dr Tommaso Barbi, for guidance in molecular biology; and Sofie Vonlanthen, for being there for me from start to finish. I would also like to thank Dr Chloe Economou, Dr Janet Waterhouse, Dr Rosie Young, Alice Lei, and Laura Stoffels for assistance with proof reading, and Khai Kong Jien for help with generation of raw bioinformatics data.

Finally I would like to thank my family for their assistance, and most of all Ria Moody; for sharing in my excitement when things went well, providing encouragement when they didn't, and being utterly supportive throughout the entire process.

Abstract

The *Chlamydomonas reinhardtii* chloroplast represents an attractive platform for therapeutic protein production, not least because of the availability of routine techniques for foreign gene expression, the low cost of cultivation, and the lack of endotoxins or potentially infectious agents in the algal host. As an application of these techniques, the primary focus of this thesis has been the expression of bacteriophage endolysins in the *C. reinhardtii* chloroplast. Endolysins hold great promise as antibacterials since they can bring about the lysis of a particular bacterial pathogen without affecting the body's natural flora, do not result in acquired resistance in the pathogen, and can kill pathogens that colonize mucosal surfaces and biofilms.

The expression of the lysin *cpl-1* specific to the major human pathogen *Streptococcus pneumoniae* has been confirmed in the *C. reinhardtii* chloroplast. The enzyme has subsequently been purified, and its lytic activity against culture collection and clinical strains of *S. pneumoniae* has been demonstrated, both for crude and enriched extracts.

Two further endolysins, *gp20* (specific to *Propionibacterium acnes*, strongly associated with clinical acne vulgaris) and *lys16* (specific to *Staphylococcus aureus*, a common hospital acquired infection), have failed to express to detectable levels. This has instigated new investigations into the various factors affecting foreign gene expression in the *C. reinhardtii* chloroplast. Research has been conducted both in a wet lab context (with the use of modified leader sequences, full protein fusions and overhauled gene design) and *in silico* (looking particularly at codon- and codon pair usage in a wide panel of endogenous and recombinant genes).

A defined codon pair bias has been shown to be present in the *C. reinhardtii* chloroplast, the first such bias to be reported in any organelle. The codon preferences observed have been related to a panel of transgenes that have previously been introduced into the chloroplast in the Purton lab, although no correlation has been found between codon pair usage and transgene expression.

Abbreviations

AMPS	ammonium persulphate
BAC	bacterial artificial chromosome
BHI	brain heart infusion
BSA	bovine serum albumin
CAI	codon adaptation index
CBD	cell binding domain
CD	catalytic domain
CES	control by epistasis of synthesis
ChlA	chlorophyll A
ChlB	chlorophyll B
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	2'deoxynucleoside 5'-triphosphate
DTT	dithiothreitol
DW	dry weight
ECL	enhanced chemiluminescence
EDTA	ethylenediaminetetraacetic acid (disodium salt)
gDNA	genomic deoxyribonucleic acid
hGH	human growth hormone
HSM	high salt minimal medium
IgG	immunoglobulin G
LB	luria-bertani medium
mRNA	messenger ribonucleic acid
NCBI	National Centre for Biotechnology Information
OD	optical density
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PG	peptidoglycan
PSI	pounds per square inch

rDNA	ribosomal deoxyribonucleic acid
RNA	ribonucleic acid
RNAP	ribonucleic acid polymerase
rRNA	ribosomal ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
STGG	skimmed milk tryptone glycerol glucose medium
TAP	tris acetate phosphate medium
TBS	tris buffered saline
TBS-T	tris buffered saline – tween 20
TE	tris EDTA
TEMED	N, N, N', N'-tetramethylethylenediamine
tris	tris (hydroxymethyl) aminomethane
tRNA	transfer ribonucleic acid
TSP	total soluble protein
TSY	trypticase soy yeast extract medium
UTR	untranslated region
v/v	volume for volume
w/v	weight for volume
WT	wild type

Table of contents

Chapter 1

Introduction

1.1 The endolysins, an overview

1.1.1 The limitations of current antibiotics

In recent years the limitations of our antibiotic arsenal have become increasingly apparent. The most publicised (and arguably most important) issue is that of antibiotic resistance and the emergence of multidrug resistant 'super bugs'. There are, however, several other factors that make traditional antibiotics less than ideal, including issues of low specificity for bacterial pathogens and the shortage of antimicrobial agents capable of targeting mucosal membranes and biofilms.

With a few exceptions, traditional antibiotics tend to be broad or medium spectrum in their antibacterial activity; this can be considered both a success and a failure of modern medicine. The issue of specificity is very much a double-edged sword. On the one hand, broad-spectrum antibiotics are incredibly versatile making them convenient in their usage and cost-effective in the scale by which they can be licensed and produced. On the other hand, the disruption of the natural flora of commensal bacteria can leave patients open to fungal and protozoan infections, especially in those individuals with weakened immune systems (Huppert *et al.*, 1953). Recently this has become a particular issue as disruption of the natural flora also paves the way for opportunistic infection by antibiotic resistant bacteria (Donskey, 2006).

The lack of agents capable of targeting mucosal membranes and biofilms is another shortcoming of conventional antibiotics. With the exception of mupirocin and polysporin, mainstream antibiotics are unable to act on bacteria colonizing the mucosal membranes (Hudson, 1994). This is significant as mucosal membranes have been shown to be a common start point for human infection, and often act as a reservoir for pathogenic bacteria (von Eiff *et al.*, 2001). In an effort to curb the spread of multi-resistant pathogens such as methicillin resistant *Staphylococcus aureus* (MRSA), such mucosal membrane targeting antibiotics are routinely prescribed prophylactically in hospitals and other high-risk environments, with some degree of success (Hudson, 1994; Perl *et al.*, 2002). Unsurprisingly, the emergence of resistant strains is already being observed, for example the several distinct lines of *S. aureus* that now exhibit mupirocin resistance (Cookson, 1998).

1.1.2 A brief history of antibiotic resistance

The issue of antibiotic resistance is possibly the most pressing of all the concerns relating to the use conventional antibiotics, with some members of the scientific community predicting a return to the pre-antibiotic era (Davies and Davies, 2010). The factors that contribute to the current shortage of effective antibiotics are many, but can essentially be condensed down to two: the emergence and rapid spread of antibiotic resistance (at least in part due to misuse), and a decrease in the development of novel classes of antibiotics due to Big Pharma reluctance to invest in new discovery programmes, and depletion of the 'low hanging fruit' (Payne *et al.*, 2007). It is important to appreciate that the development of antibiotic resistance in bacterial species is not a new phenomenon – the process itself predates mankind – rather, it is the stagnation of novel antibiotic development that has made us so acutely aware of it. The decline in antibiotic discovery is summarised in Figure 1.1, and excellently reviewed by Davies and Davies (2010).

In order to understand the issue of resistance development and spread, it is also important to consider the various means by which resistance can develop. For the purpose of this review, the β-lactams shall be used as a case study. Since the discovery of the first β-lactam antibiotic, penicillin, in 1928 we have been locked in an arms race with the bacterial world. As each new antibiotic is introduced, resistance rapidly develops and spreads. This requires new antibiotics to be developed to take their place giving but a short grace period before the cycle begins again (Llarrull *et al.*, 2010). The first appearance of β-lactam resistance was seen in 1940 by members of the penicillin discovery team, and took the form of a β-lactamase (Abraham and Chain, 1940). The β-lactamases are a family of enzymes that confer resistance by the first of the three major mechanisms of resistance to be discussed here – antibiotic deactivation. The β-lactamases are capable of cleaving the β-lactam square ring required for antibiotic activity, thus providing protection for the bacterium. The next generation of β-lactams (comprising methicillin and oxacillin) addressed this by means of a semi-synthetic modified β-lactam ring that was resistant to attack from β-lactamases. Alongside these new β-lactamase resistant β-lactams came a range of direct β-lactamase inhibitors including clavulanate and the penicillin sulfones (Drawz and Bonomo, 2010).

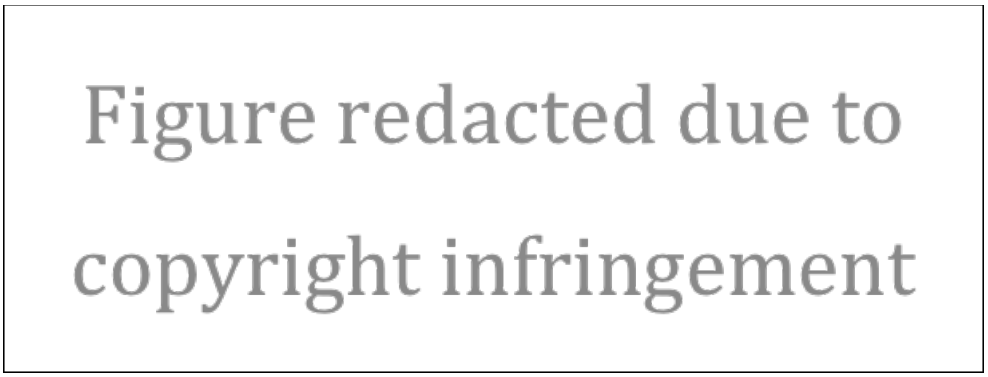


Figure 1.1 – A timeline of antibiotic development over the past 70 years

The various stages in antibiotic development, and the corresponding development of resistance. The majority of the antibiotic classes used today were developed during the Golden age. The last 50 years have been characterised by various less successful approaches to novel antibiotic discovery culminating in the disenchantment of the large pharmaceuticals following years of minimal return on major investments. Reproduced from (Davies and Davies, 2010).

These were used both to extend the life of previous generation β-lactams, and also concurrently with the next generation β-lactams in an effort to pre-empt future resistance. New resistant strains were quick to emerge however, with MRSA at the forefront. This time a different resistance mode was employed by the bacteria - target site modification. In this mechanism the antibiotic's target is either altered so that it is no longer inhibited by the antibiotic, or bypassed completely so any inhibition is no longer deleterious. In the case of the β-lactam target, the penicillin binding protein (PBP) was exchanged with a related, but structurally different PBP, PBP-2a, which showed a greatly diminished affinity to the β-lactams (Antignac and Tomasz, 2009). Bacteria with PBP-2a were able to circumvent both the new β-lactams and the β-lactamase inhibitors, effectively taking antibiotic development back to square one (Lowy, 2003). The β-lactams have here illustrated two of the three major antibiotic resistance mechanisms, the third being the restriction of antibiotic accumulation within the cell. This can be mediated either by reduced permeability of the cell, or by direct efflux of the antibiotic. The latter is commonly seen in Gram-negative bacteria and is particularly effective in that a single pump can typically efflux several different antibiotics. Additionally, genes encoding the efflux machinery are commonly seen on mobile elements such as plasmids or transposons and are therefore easily spread by horizontal gene transfer (Nikaido, 1998). The three resistance mechanisms discussed are summarised in Figure 1.2.

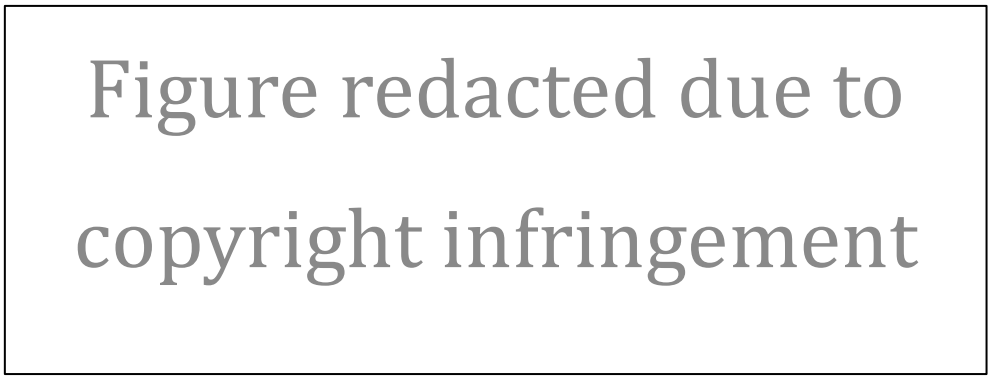


Figure 1.2 - A schematic illustration of the three major antibiotic resistance mechanisms

The main resistance mechanisms are illustrated as seen for the β-lactams: **a1)** reduced accumulation via cell permeability, **a2)** reduced accumulation via drug efflux, **b)** enzymatic deactivation (β-lactamase activity), and **c)** target site modification (PBP*). Reproduced from (Llarrull *et al.*, 2010).

1.1.3 The re-emergence of bacteriophage therapy

The '90s and early '00s saw the rise of structural and genomic based rational drug design and high throughput screens. Despite heavy investment, significant returns were not seen (Payne *et al.*, 2007), prompting a gradual shift back to the search for novel antibiotics from natural systems (Luzhetskyy *et al.*, 2007). One such avenue which has been re-opened for investigation is that of phage therapy (Fischetti, 2001; Housby and Mann, 2009). Bacteriophages (or simply 'phages') are viruses that specifically target bacteria and can be considered the most abundant biological entities on the planet, numbering an estimated 10^{31} particles (O'Flaherty *et al.*, 2009). They are generally regarded as harmless to human cells but are very effective in killing bacteria, and their capacity for auto-amplification *in vivo* allows high efficacy from a single dose. Importantly, they show high specificity for their target bacterium and thus do not disturb the natural bacterial flora. In addition the co-evolved nature of a bacterium and its corresponding phage imply low occurrence of resistance (Kutateladze and Adamia, 2010).

Bacteriophage particles were discovered by Felix d'Herelle in 1917, and investigations into their potential as therapeutic agents were instigated as early as 1919 (Housby and Mann, 2009). Phage therapy became popular in the 1930s, but despite some success, efficacy was far from reliable. This was commented on in 1934 in a report by the American Medical Association's Council on Pharmacy and Chemistry where the case was made that the present understanding of phage particles was severely limited (Eaton and Bayne-Jones, 1934). Looking back this was indeed the case; one preparation of phage particles listed phenol as one of its 'preservatives', which would have almost certainly deactivated any phage present (Housby and Mann, 2009). Interest in phage therapy in the Western world rapidly waned with the rise of antibiotics and such treatments were rarely seen after the late 1940s. Phage therapy did however continue behind the Iron Curtain, and is still in use in several former Soviet block countries. An example of such continued use of phage particles to treat infection can be seen in the Eliava Institute in Georgia, which was at one point the site of phage particle production for the entire Soviet Union (Kutateladze and Adamia, 2010). The institute cites many successful cases of treatment but as of yet there have been no large-scale clinical trials showing sufficient scientific rigor (such as randomization and double-blind trials)

to satisfy Western authorities (O'Flaherty *et al.*, 2009). Despite this, several phage regimes have been given FDA approval for use as food preservatives for human consumption, indicating that therapeutic use should not be ruled out (Housby and Mann, 2009). There are, of course, also disadvantages associated with phage therapy. These include rapid inactivation in the spleen, immune clearance, decreased efficacy with repeat exposure, and inherent problems related to bacterial endotoxin contamination (Fischetti, 2001). There are also concerns over the rapid evolution of the virus particles *in vivo* leading to unpredictable new phenotypes. The volatility of using a complete biological entity as a therapeutic has led to a reductive approach to the issue, with phage-encoded peptidoglycan hydrolases providing a promising alternative.

1.1.4 Recombinant bacteriophage endolysins as novel therapeutics

Characterization of the molecular mechanisms of the phage lifecycle has revealed a family of peptidoglycan hydrolases known as endolysins, commonly abbreviated to 'lysins'. These enzymes have the potential to take the place of whole phage particle based treatment of bacterial infection (Fischetti, 2008).

1.1.4.1 Lysins in the native phage lifecycle

Endolysins are bacteriophage-derived enzymes that perform an essential role in the release of progeny from the bacterial host cell (Figure 1.3) (Llarrull *et al.*, 2010). They are generally expressed early in the infection cycle and accumulate in the bacterial cytoplasm until a genetically specified time when they are allowed through the cytoplasmic membrane via the oligomerisation of pore-forming proteins known as holins (Wang *et al.*, 2000; Ziedaite *et al.*, 2005). On exposure to the peptidoglycan cell wall, the lysins degrade the structure by one of several modes of action targeting either the peptide or glycan bonds (Figure 1.4), and resulting in bacterial cell lysis and release of viral particles into the environment (Young, 2000). Lysins that target Gram-positive cell walls have been shown to display incredibly tight binding to their substrate with nanomolar affinities being observed. This is presumably to prevent the diffusion of lysins resulting in the destruction of nearby uninfected cells that could otherwise serve as new hosts. Such high binding affinities are not observed for Gram-negative lysins where local cells are protected by their outer membrane (Loessner *et al.*, 2002).

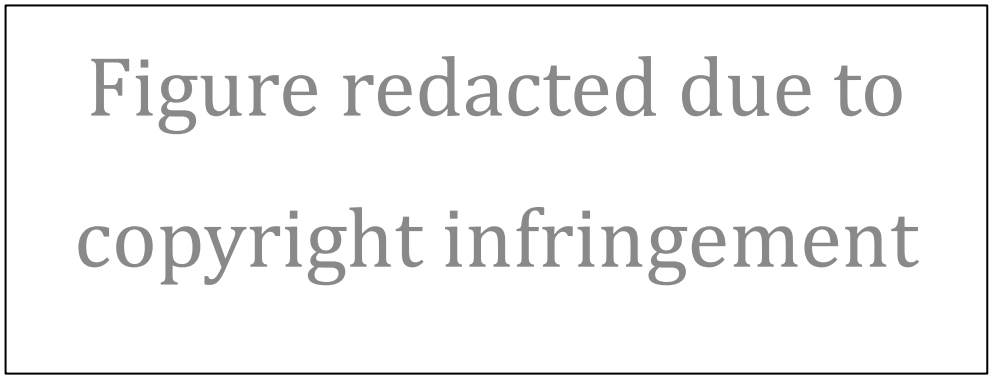


Figure 1.3 – The bacteriophage lytic life cycle

The bacteriophage lytic life cycle begins with the binding of the phage to the target bacterium **(1)**. Nucleic acid is injected into the cell **(2)**, where it proceeds to hijack cellular function and initiate degradation of host DNA **(3)**. Phage particles are then synthesised, along with lysins for progeny release **(4, 5)**. At a genetically specified point, pore forming holin proteins allow the lysins through the cytoplasmic membrane resulting in cell lysis and progeny escape **(6)**. Many Gram positive lysins have been shown to remain bound to their substrate post-lysis to prevent destruction of neighbouring uninfected cells. Figure adapted from www.hyglos.de.

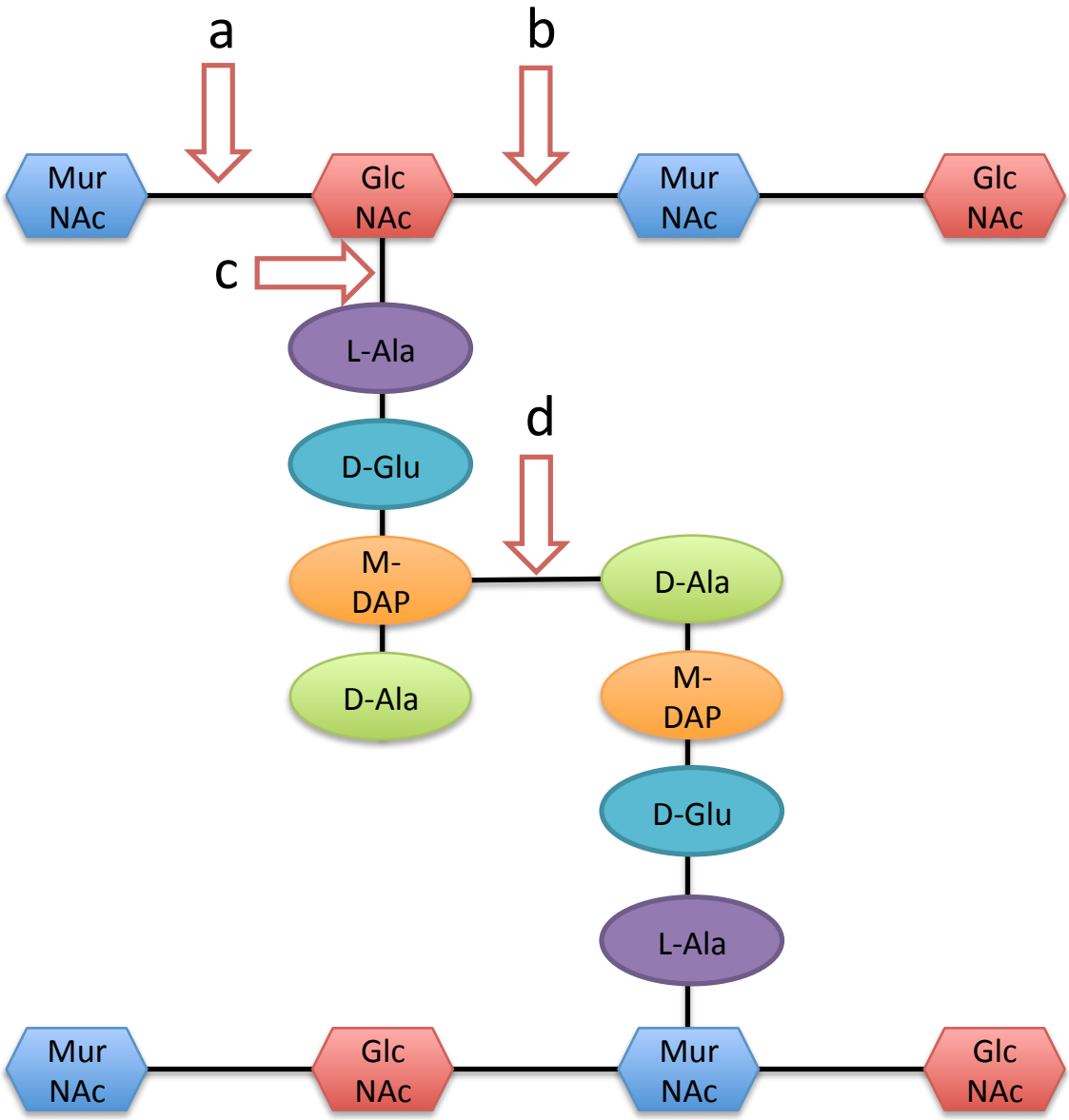


Figure 1.4 – Modes of action commonly displayed by phage lysins

Lysins show a variety of peptidoglycan hydrolase mechanisms targeting glycan, amide, and peptide bonds. A selection of common modes of action is displayed: **a)** *N*-acetyl-β-D-muramidase **b)** *N*-acetyl-β-D-glucosaminidase **c)** *N*-acetylmuramoyl-L-alanine amidase **d)** Endopeptidase.

1.1.4.2 ***Lysins as therapeutics***

It has been shown (Loeffler *et al.*, 2001; Rashel *et al.*, 2007) that lysins can be synthesised using recombinant platforms, remain active *in vitro* and *in vivo*, and are fully capable of working exo-lytically on Gram positive bacteria. (That is to say, digesting the cell wall from the outside as an antimicrobial agent, as opposed to endo-lytically as per the native function; a process referred to as 'lysis from without' relative to 'lysis from within'.) Lysins have been identified that target a number of Gram-positive pathogens, a selection of which is listed in Table 1.1. Potential uses for lysins extend past their direct application as primary therapeutics to areas such as food preservation and as a defence against bio-warfare.

The external nature of these agents automatically lends an advantage over traditional antibiotics, which tend to require cell internalization for activity. By never entering the bacterial cell many of the common resistance mechanisms discussed above are circumvented: reduction of accumulation is obviously not an issue, and antibiotic deactivation is generally effected within the cell and thus extra cellular deactivation would require the efflux of large amounts of a deactivation mediator at high metabolic cost. This leaves only target site modification or bypass. The highly integrated life cycles of bacteria and phage, however, combined with the absolute necessity of a functional lysin for phage genetic material to persist, seem to have resulted in lysins that target essential components of the cell wall. The co-evolutionary process has been likened to a biological high throughput analysis of the host bacterium (Fischetti, 2008), to the extent that a recent study has used a 'lysin identified' cell wall component as a target for a small molecule inhibitor antibiotic (Schuch *et al.*, 2013).

Table 1.1 – A selection of lysins targeting relevant bacterial pathogens

Lysin	Phage	Bacterial target	Catalytic activity	Mass (kDa)	Potential application	Reference
Cpl-1	Cp-1	*Streptococcus pneumoniae*	Muramidase	39	Pharmaceutical therapeutic	(Loeffler *et al.*, 2001)
LysK	Phage K	*Staphylococcus aureus*	Endopeptidase & amidase	54	Pharmaceutical therapeutic	(O'Flaherty *et al.*, 2005)
CD27L	ΦCD27	*Clostridium difficile*	Amidase	32	Pharmaceutical therapeutic	(Mayer *et al.*, 2008)
PlyPSA	PSA	*Listeria monocytogenes*	Amidase	35.3	Food preservation/ surface decontamination	(Zimmer *et al.*, 2003)
PlyL	λ prophage Ba02	*Bacillus anthracis*	Amidase	26.4	Bio-warfare defence	(Low *et al.*, 2005)

To date, no native endolysin resistance has been reported. Direct attempts to generate resistant cell lines using low dose lysins on both solid and liquid media have failed to give resistance even after 40+ cycles (Loeffler *et al.*, 2001). This should not of course be interpreted as evidence that resistance is not possible. Clearly any such parasite:host co-evolution will result in developments on both sides of the metaphorical divide, and bacteria have indeed evolved numerous antiviral strategies (Labrie *et al.*, 2010). It is conceivable however that endolysin resistance does not currently feature in this arsenal simply because at the point of cell lysis the bacterial host cell will generally already be dead. Although there are documented processes for utilitarian defence at the cost of the individual cell (known as abortive infection systems), these all act before the phage particles are completely assembled. Additionally, due to the irreversible nature of binding seen in many Gram-positive lysins, it is likely that throughout their evolutionary history, live bacterial target cells have very rarely come into external contact with lysins. Any resistance would therefore have to be developed *de novo*, as opposed to simply being acquired via horizontal gene transfer from resistance already present in the environment. This will hopefully extend the resistance-free window of a lysin therapeutic.

Lysin target ranges are in general very narrow, reflecting the highly specific nature of the host bacteriophages for the target bacterium. Lysin activity tends to be restricted to a single organism, with many showing species, serotype, and even strain specificity (Fischetti, 2003). This can be of great benefit for clinical applications as pathogenic organisms can be targeted without disruption of the natural flora. This might also delay the emergence of resistant bacteria, as a selective pressure will only be placed on the pathogenic target, rather than the entire bacterial community. Furthermore, lysins have been shown to be active on mucosal membranes, as evidenced by bacterial clearance of *in vivo* colonisation models (Nelson *et al.*, 2001). They have also been shown to be effective in the disruption and clearance of biofilm structures (Sass and Bierbaum, 2007).

The specific cell wall degrading properties of the lysins have been shown to give high degrees of synergy with conventional antibiotics such as penicillin and gentamicin (Djurkovic *et al.*, 2005), and also with lysins displaying different

catalytic modes of action. It has been demonstrated *in vivo* that when used simultaneously, the lysins Cpl-1 and Pal administered at 2.5 μg each can give a similar treatment outcome to 200 μg of Cpl-1 or Pal as applied separately (Jado *et al.*, 2003; Loeffler and Fischetti, 2003).

1.1.4.3 ***Issues with therapeutic use of lysins***

Lysins can be seen to address many of the problems observed for conventional antibiotics, such as resistance development, specificity, and activity on mucosal membranes and biofilms. They do however bring with them a new set of complications. A major constraint to the use of lysins as therapeutics is the necessity for the peptidoglycan to be accessible. This permits lysis of the Gram-positive cell wall where the peptidoglycan is directly accessible to the surroundings; the outer membrane of Gram-negative species prevents such access and thus blocks antimicrobial activity. There are isolated cases where lysins are shown to be active on Gram-negative strains due to intrinsic membrane disrupting properties of the lysin concerned, and there is much interest in recombinantly reengineered lysins with Gram-negative activity, but at present the therapeutic use of lysins can be seen to be limited to Gram-positive pathogens (Schmelcher *et al.*, 2012).

Even accepting this limitation, many more issues with lysins as therapeutics can be seen to stem from their proteinaceous nature. Even human protein therapeutics such as humanised insulin face difficulties relating to drug delivery and bioavailability. With non-native proteins there are additional complications relating to rapid degradation by the patient's immune system, the generation of deactivating antibodies, and the risk of severe immunogenic reactions. The latter is compounded by the release of highly immunogenic bacterial cell components during the lytic process. Bioavailability is also a concern due to the size of the molecules involved. Oral delivery is not suitable due to degradation in the stomach followed by issues involving traversing the gut epithelium. Even with direct administration via intravenous or intraperitoneal injection, access to the target bacteria can still be a problem, notably in the cases of intracellular pathogens such as *Chlamydia ssp.* and granulomatous inflammatory diseases such as tuberculosis.

It has been demonstrated however, that many of these problems are surmountable. Several *in vivo* studies have shown success with intravenous (Loeffler *et al.*, 2003), intraperitoneal (Witzenrath *et al.*, 2009) and nasal (Loeffler *et al.*, 2001) administration of lysins in mice. The effective clearance of infection in these cases despite immune degradation has been attributed to the rapid action of the lysin on the target bacterium balancing the effect of short serum half-lives (20 minutes in the case of the Cpl-1 lysin) (Entenza *et al.*, 2005). This is reinforced by the extremely fast action of various lysins demonstrated *in vitro*, with log death of bacterial specimens occurring in the first few seconds after application (Nelson *et al.*, 2001).

Immunogenic studies into Cpl-1 from the *Streptococcus pneumoniae* phage Cp1 have shown it to be immunogenic, however enzyme activity was not significantly reduced by hyperimmune rabbit serum. Studies on mice exposed to Cpl-1 four weeks prior to treatment tested positive for IgG against Cpl-1, but treatment outcome was not significantly different to untreated mice (Loeffler *et al.*, 2003). It has been suggested that the rapid activity and extreme binding of many lysins allow them to 'out-manoeuvre' the immune system *in vivo* (Schmelcher *et al.*, 2012). Further investigation has shown that repeated exposure of Cpl-1 does not induce any side effects indicative of a hypersensitivity reaction such as anaphylaxis (Jado *et al.*, 2003).

Although activity is observed *in vivo* even with rapid immune clearance, various attempts have been made to reduce the immunogenicity of lysins, again with Cpl-1 being used as a model system. Cysteine specific PEGylation, a proven immune masking tool, has shown to abolish antibacterial activity for Cpl-1 (Resch *et al.*, 2011a); however, promising results have come from dimerisation studies via introduction of C-terminal cysteine residues to allow disulphide bond formation. In the latter case, activity was also boosted in addition to the desired drop in plasma clearance rates (Resch *et al.*, 2011b). Targeting of intracellular bacteria is still an issue, however recent work has suggested one solution to be the fusion of lysins with cell penetrating peptides, thus allowing transduction into affected cells (Borysowski and Gorski, 2010).

1.1.4.4 ***Lysin structure and function***

Gram-positive lysins tend to have a discrete modular structure with separate catalytic and cell binding domains (CD and CBD respectively), commonly arranged as shown in Figure 1.5. On a rudimentary level it can be said that the CBD is responsible for lysin specificity, whereas the CD is purely liable for the cleavage of the peptidoglycan, and indeed this division has been demonstrated in early domain swapping experiments (Díaz *et al.*, 1990). However, it has also been shown that in many cases the CBD is required for catalysis, and that the CD can contribute to the specificity observed. The rationale for both such phenomena can be found in the interplay between the two domains. The few crystal structures of whole lysins that are available (crystallisation often being hindered by the flexibility of the linker region joining the two domains) have suggested that the CBD is important for positioning the catalytic domain relative to the substrate (Hermoso *et al.*, 2003), in a way that is not unlike what has been documented for other glycan hydrolases such as certain cellulase enzymes (Bolam *et al.*, 1998). CDs do not display the same level of cell wall specificity seen for CBDs, but are still bond specific, so their activity will depend on this bond being present within the peptidoglycan structure.

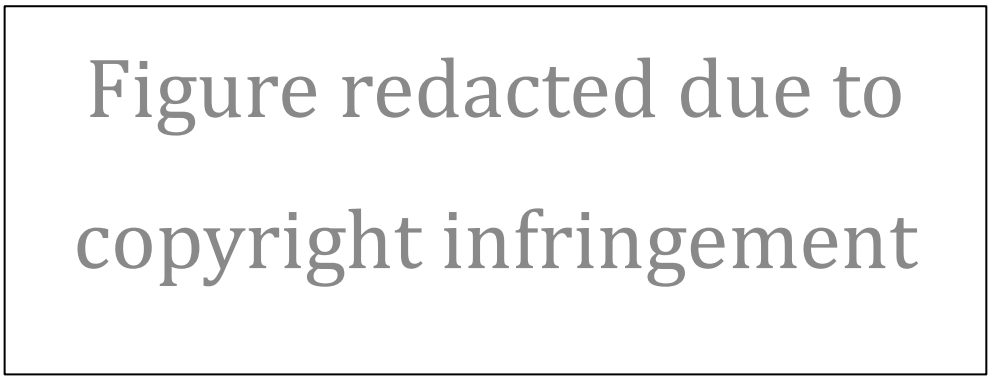


Figure 1.5 – A general schematic showing the modular nature of many Gram-positive lysins (a), with the crystal structure of the PlyPSA lysin (b)

The cell-binding domain is generally thought to confer specificity while the catalytic domain is responsible for cleavage of the peptidoglycan cell wall; however, interplay between domains has been shown to be important in several cases. **(a)** reproduced from (Fischetti, 2008), **(b)** reproduced from (Korndörfer *et al.*, 2006).

Lysins that deviate from the standard two domain structure are also seen. Many Gram-negative lysins consist solely of a CD, having less need for tight binding and hence specificity as discussed above. Other Gram-positive lysins have multiple CDs, although studies have shown these not to be equal in terms of catalytic activity. A selection of different lysin structures is shown in Figure 1.6 reproduced from (Schmelcher *et al.*, 2012). In recent years there has been a considerable interest in the mixing-and-matching of domains to produce 'custom lysins', as well as reductive studies investigating the minimum requirements for catalytic activity (Entenza *et al.*, 2005). It has been noted that in many cases the CD domain shows significant homology to the host bacterium's own autolysin; however, there are also cases where this is not the case. Furthermore there are lysins where the CD shows clear homology to peptidoglycan hydrolases from bacteria or phage species completely distinct from the host organism, forming natural chimeric endolysins. An example is the Pal lysin, which has a CBD that shows homology to other *S. pneumoniae* phage lysins; however, it has a CD that shows homology to the lysin of the *Lactococcus lactis* phage BK5-T (Sheehan *et al.*, 1997).

1.1.4.5 ***Issues with current production platforms***

As with many protein-based therapies, the specificity and efficiency of the lysins is due to their complexity, which in turn requires complex biological systems for their production. Such production platforms generally result in a far higher cost than their small molecule counterparts (Dove, 2002). The particular issue with the production of lysins is that, unlike other protein therapeutics such as insulin, they are in direct competition with small molecule equivalents. Despite the potential advantages over the conventional antibiotics, the cost of production will be a major hurdle in terms of bringing such a product to market. Currently used production systems for lysins have been almost exclusively bacterial, although high level expression has been reported in tobacco (Oey *et al.*, 2009a). Neither of these platforms is ideal for lysin production as discussed below.

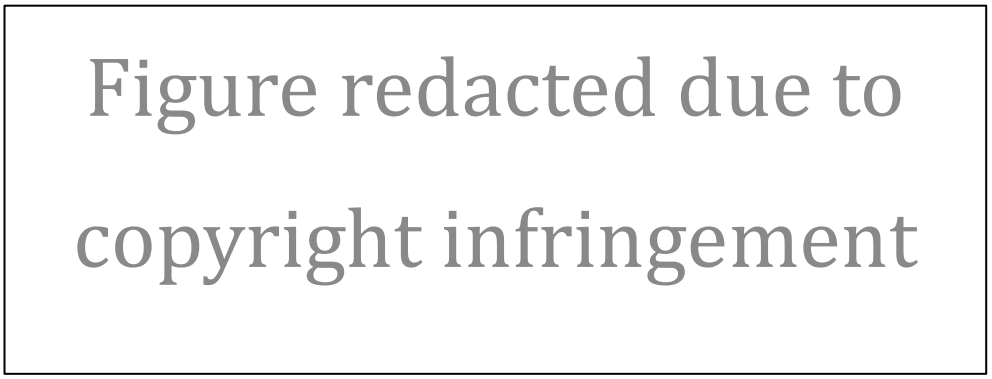


Figure 1.6 - Schematic representation of a selection of lysins showing traditional, three domain, and inverted architectures

Reproduced from (Schmelcher *et al.*, 2012).

1.2 Current and future platforms for therapeutic protein synthesis

1.2.1 The rise of biologics

In recent decades the fields of molecular biology and biotechnology have developed rapidly, opening the doors for an entirely new domain of pharmaceuticals know as biologics. This new range of therapeutics are characterised by their necessity to be produced in biological systems and include enzymes, hormones, antibodies and vaccines. They offer many advantages over conventional small molecule drugs such as increased efficacy, high specificity (greatly reducing side effects), and the ability to access targets simply not available to other agents (Dove, 2002).

1.2.2 Current expression platforms and their limitations

With the advent of recombinant technologies came the ability to produce human therapeutic proteins such as insulin, erythropoietin, and somatropin by means of transgenic expression platforms. These were naturally a significant improvement over previous solutions (which included the harvesting of therapeutic protein from cadavers), although they did come with their own limitations. Bacterial systems are often unable to correctly fold complex eukaryotic products and contain endotoxins that require strict removal from the final product. Yeast platforms are able to carry out more complex eukaryotic protein folding and are free from bacterial toxins, but have issues with hyper- or inappropriate glycosylation. Mammalian cell production systems such as Chinese Hamster Ovary (CHO) cells are expensive, difficult to scale up, and carry the danger of latent contamination with viral or prion particles (Rasala *et al.*, 2010). Ultimately however, it is the cost of protein based therapies that is seen as the major obstacle to their widespread adoption, with base production accounting for up to 25 % of the final retail cost of a therapeutic (Corchero *et al.*, 2013). Of this 25 %, up to 80 % can be due to the downstream processing of the product, making this an area of particular interest for development (Roque *et al.*, 2004).

Since the production of recombinant human somatostatin in 1977 (Itakura *et al.*, 1977), followed by insulin in 1979 (Goeddel *et al.*, 1979), there have been significant improvements in recombinant protein expression platforms, ranging

from boosts in productivity and reduction of protein aggregation, to the development of more humanised post translational modification systems. There is a faction of the academic community however who feel current production platforms are reaching their physiological limits, and therefore non-conventional platforms should be considered for future investigation (Corchero *et al.*, 2013).

1.2.3 Recombinant plants as an alternative platform

The use of plant cells for the production of therapeutic proteins addresses many of the problems associated with conventional platforms, and has recently been gaining attention under the umbrella term of 'molecular farming'. The eukaryotic nature of plant cells allows for complex product formation; they are devoid of the endotoxins seen in bacterial systems, and are readily scaled up to industrial scale dramatically reducing costs (Ma *et al.*, 2005). Additionally, they have fewer issues with inappropriate glycosylation relative to yeast production, and do not pose the danger of passing on mammalian pathogens or prion particles. Recombinant plant technologies have had a slower start than their microbiological or mammalian counterparts, but since the production of human serum albumin in tobacco in 1990 (Sijmons *et al.*, 1990), they have been rapidly gaining attention as a next generation expression platform.

As with any system however, molecular farming is not without its limitations. Concern has been raised as to the release of transgenic material into the environment; this is particularly true for genes encoding molecules displaying human bioactivity owing to the danger of recombination with food crops (Dove, 2002; Mayfield *et al.*, 2007). The use of chloroplast transformation has greatly reduced the risk of this occurring as the chloroplast genome is maternally inherited in most crop species, and thus is generally not present in pollen. Despite this, paternal inheritance of chloroplasts has been shown to occur, albeit at very low frequencies (Wang *et al.*, 2004). The concept of open field systems for therapeutic protein expression has also been met with public opposition owing to the possibility of contamination acting in the opposite direction, with the introduction of environmental toxins and pathogens into the final product (Fischer *et al.*, 2012). Aside from containment issues, higher plants have lengthy generation times for the creation of recombinant lines, and slow accumulation of biomass

relative to other platforms (Corchero *et al.*, 2013). Tobacco for example can take up to two years to go from initial transformation to commercial quantities of product (Manuell *et al.*, 2007). Expression of biologics in the green unicellular alga *Chlamydomonas reinhardtii* can address many of these issues.

1.3 *Chlamydomonas reinhardtii* as an expression platform

1.3.1 An introduction to *Chlamydomonas reinhardtii*

C. reinhardtii is a unicellular freshwater micro-alga with a diameter of ~10 μm, and a doubling time of around eight hours. Figure 1.7 shows a false colour image of a wild type *C. reinhardtii* cell, and a schematic showing the major physiological features. It has a single chloroplast that occupies a large portion of the cell volume (reports range from 40-80 %), and contains between 80 and 100 copies of the plastid genome. Crucially *C. reinhardtii* represents an interesting bridge-point between the plant and animal kingdoms; it carries out higher plant-like photosynthesis, yet is also motile by virtue of its two animal-like flagella. As a result of its interesting biology, genetic tractability, and ease of cultivation it has been used as a model organism for over 50 years. Research has been conducted in many distinct fields including flagella assembly and function (Ringo, 1967), phototaxis (Foster *et al.*, 1984; Nagel *et al.*, 2003), circadian rhythms (Suzuki and Johnson, 2001), and the photosynthetic electron transport chain (Rochaix, 2011). Although autotrophic, *C. reinhardtii* is capable (given a suitable carbon source) of growing heterotrophically in the absence of light, or even in the event of the disruption of its photosynthetic machinery (Harris, 1989). The cell is encapsulated in a proteinaceous cell wall comprising of up to seven discrete layers (Roberts *et al.*, 1972). Cell wall deficient mutants are viable, and several such mutants (e.g. cw15) are commonly used in basic and applied research (Hyams and Davies, 1972).

C. reinhardtii reproduction can occur both sexually and asexually. When in the vegetative state cells are haploid, fall into one of two genetic mating types (plus or minus), and reproduce by binary fission. On entering the sexual cycle, (typically triggered by nitrogen deprivation or by other abiotic stress conditions), cells become sexually competent gametes of either plus or minus mating type and fuse

with their opposite mating type. Fusion is initiated by the intertwining of flagella which triggers the release of the cell wall via secretion of the metalloprotease, autolysin (Buchanan *et al.*, 1989; Kinoshita *et al.*, 1992). The diploid zygote then forms a dormant structure known as a zygospore that is capable of enduring adverse environmental conditions. On re-establishment of favourable growth conditions, the zygospore germinates, dividing by meiosis to form four haploid daughter cells – two mating type plus and two mating type minus. During sexual reproduction the chloroplast is inherited from the mating type plus parent, the mitochondria from the mating type minus, and the nuclear genome is inherited in a Mendelian manner (Harris, 1989).

In light of the recent interest in microalgae as a source of renewable energy, there has been considerable interest in *C. reinhardtii* for the production of biofuels in the form of bio-diesel (derived from storage lipids) and bio-hydrogen (derived from excess reducing equivalents channelled to protons by the enzyme hydrogenase). However, owing to the relatively low productivity in this species, research has now mostly shifted to the use of *C. reinhardtii* as a genetic model for other more productive algal species (Rupprecht, 2009). More recently still, interest in bulk commodities from microalgae has declined, and focus is moving increasingly towards utilisation of *C. reinhardtii* for higher value products ranging from specialised 'designer fuels', to industrially relevant compounds such as the diterpenoids (Lohr *et al.*, 2012), and therapeutic proteins (Mayfield *et al.*, 2007).

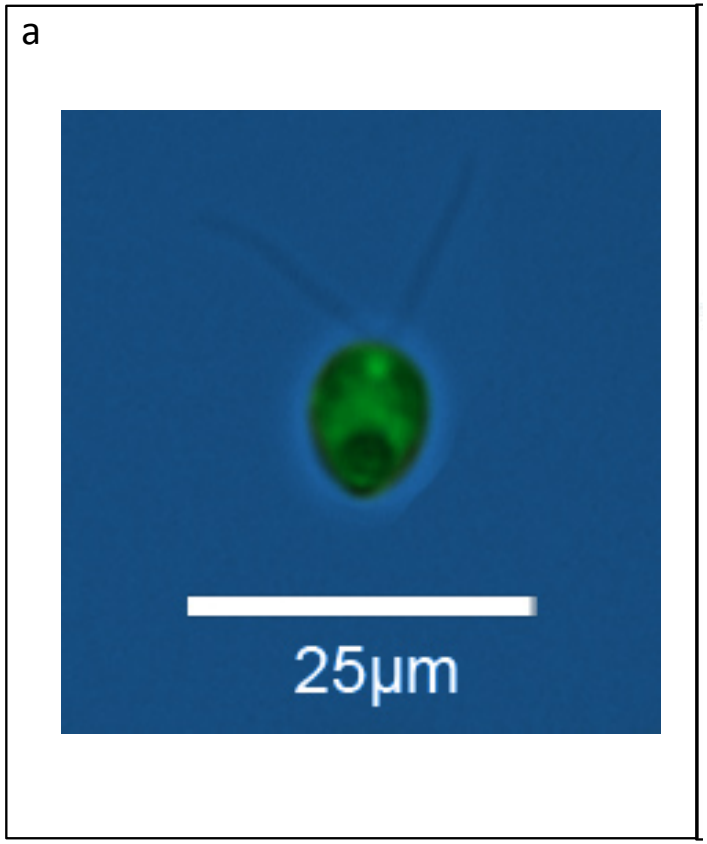


Figure redacted due to copyright infringement

Figure 1.7 – False colour image of *C. reinhardtii* (a) and a schematic representation showing the major physiological features of the cell (b).

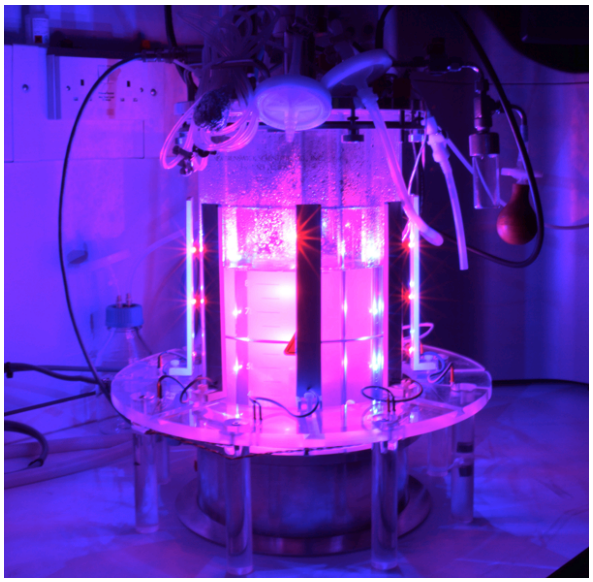
The image shown in panel **(a)** was captured using an EVOS FL digital light microscope (PEQLAB) and false coloured using Adobe Photoshop CS5. The diagram in panel **(b)** is reproduced from (Merchant *et al.*, 2007).

1.3.2 Advantages of *C. reinhardtii* over higher plants

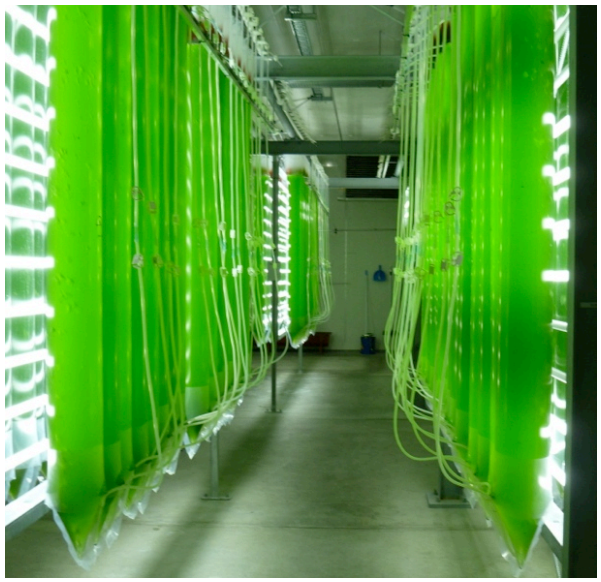
The use of *C. reinhardtii* as an expression platform for therapeutic proteins is fundamentally very similar to that of higher plants, although micro-algal transgene expression is seen to hold several key advantages over their more complex cousins. The first of these is time. As a unicellular organism that primarily reproduces asexually, developmental time is greatly reduced in relation to that of higher plant systems. This impacts all levels of production: initial transformation, where stable transgenic cell lines can be achieved in as little as three weeks; growth and harvesting of product at a laboratory scale, typically achieved in around a week; and scaling up to industrial proportions, generally taking 4-6 weeks (Manuell *et al.*, 2007).

The second major benefit of *C. reinhardtii* as a platform is the ease of containment, with cultivation being ideally suited to closed photo-bioreactor systems (Figure 1.8). This not only controls any flow of contamination (in either direction), but also allows for a far greater degree of control over growth conditions, and thus in turn a higher level of process optimisation than is possible for open field based systems.

Finally there is productivity. As is discussed in detail below, recombinant protein yields in *C. reinhardtii* are currently lower than those seen in higher plants. *C. reinhardtii* does however, have some intrinsic biological advantages which are likely to become increasingly prominent as the field matures. As a single-celled organism, every cell in a culture will express the gene of interest. This contrasts sharply with the tissue specific expression seen in higher plants. Photosynthetic productivity has also been shown to be considerably higher in microalgae as opposed to higher plants. Combined with the high total protein content seen in algae (up to 70 % fresh mass (Passwater and Solomon, 1997)), this ultimately gives organisms such as *C. reinhardtii* a far greater end-point potential as a low cost platform for recombinant protein expression (Rosales-Mendoza *et al.*, 2012). The key advantages of *C. reinhardtii* over higher plants are summarized in Table 1.2.



10s L



100s L

1000s L

Figure 1.8 – Algal photo bioreactor (PBR) systems ranging from lab to industrial scale

Left panel: 10 L lab scale PBR

Centre panel: Array of 40 L disposable hanging-bag PBRs

Right panel: Tubular PBR in custom-built complex

Table 1.2 – The major differences between higher plant and *C. reinhardtii* based transgene expression systems

	Higher Plants	***C. reinhardtii***	**Reference**
Generation of transformant lines	Months	3-4 weeks	(Purton *et al.*, 2013)
Scale up time	Months - years	Weeks	(Specht *et al.*, 2010)
Recombinant protein yields (%TSP)	Up to 70 %	Up to 10.5 %	(Oey *et al.*, 2009a; Surzycki *et al.*, 2009)
Ease of containment	Difficult	Simple	(Mayfield *et al.*, 2007)
Risk of gene flow to the environment	High	Low	(Potvin and Zhang, 2010)
Photosynthetic efficiency	-	~ 3 fold more efficient than higher plants	(Rosales-Mendoza *et al.*, 2012)

1.3.3 C. reinhardtii as an expression platform for therapeutic proteins

A range of therapeutic proteins has been successfully expressed in the *C. reinhardtii* chloroplast, a selection of which is displayed in Table 1.3. Product yields are typically in the range of 0.1-5 % of total soluble protein (TSP) (Manuell *et al.*, 2007) with a maximum reported yield of 10.5 % (Surzycki *et al.*, 2009); however, this has yet to be independently verified. Product yields are currently much lower than those seen in the more established biologics platforms, although there is much interest in addressing this as discussed below (section 1.4).

Despite low yields, expression of biologics in *C. reinhardtii* already shows promise for commercial utilisation owing to the low cost of algal cultivation. It is estimated that transgenic algal biomass can be produced for as little as $3/kg. Even given a modest value of 2 % TSP, and a soluble protein content of 25 % total biomass, this puts final product production at a cost of $0.6/g, which is already competitive with the currently established biologics platforms (Rasala *et al.*, 2010). The reduced cost associated with production is in turn matched by potentially huge reductions in downstream processing costs relative to current platforms. As with many higher plants, *C. reinhardtii* has GRAS (Generally Recognised As Safe) status, eliminating the need for stringent removal of toxic components and possibly allowing for edible delivery systems based on crude cell extracts (Gregory *et al.*, 2013; Mayfield *et al.*, 2007). Given the large proportion of total expenditure relating to downstream processing, the potential for such rudimental processing could have a significant effect on the overall economic viability of *C. reinhardtii* as an expression platform.

Table 1.3 – A selection of therapeutic transgenes expressed in the *C. reinhardtii* chloroplast

Adapted from (Specht *et al.*, 2010)

Gene Expressed	Function	Expression level	Reference
HSV8-lsc	Human anti-herpes antibody	Detectable	(Mayfield *et al.*, 2003)
CTB- VP1	Foot and mouth disease vaccine:adjuvant fusion	3 % TSP	(Sun *et al.*, 2003)
hTRAIL	Human therapeutic	~0.67 % TSP	(Yang *et al.*, 2006)
M-SAA	Bovine mammary-associated serum amyloid	~5 %	(Manuell *et al.*, 2007)
CSFV-E2	Swine fever vaccine	~2 % TSP	(He *et al.*, 2007)
hGAD65	Human therapeutic	~0.3 % TSP	(Wang *et al.*, 2008)
VP28	White spot syndrome virus vaccine	~10.5 % TSP (unverified)	(Surzycki *et al.*, 2009)
CTB-D2	*Staphylococcus aureus* vaccine:adjuvant fusion	0.7 % TSP	(Dreesen *et al.*, 2010)
Proinsulin	Human hormone	Detectable	(Rasala *et al.*, 2010)
VEGF	Human vascular endothelial growth factor	2 % TSP	(Rasala *et al.*, 2010)
αCD22PE40	Human immuno-toxin	~0.3- 0.4 % TSP	(Tran *et al.*, 2013)

1.3.4 The *C. reinhardtii* chloroplast as a sub-cellular expression compartment

The single *C. reinhardtii* chloroplast has been demonstrated to be an attractive platform for the expression of transgenes, exhibiting far higher levels of recombinant protein than achieved through transgene expression in the nucleus (Specht *et al.*, 2010). This is believed to be partly due to the extent of polyploidy seen in the chloroplast genome (referred to as the 'plastome') resulting in a high copy number of the transgene, although there are several other factors acting at the transcriptional and post-transcriptional level that contribute to the high levels of expression. The *C. reinhardtii* nuclear genome shows strong gene silencing effects, especially when multiple copies of a gene are integrated (Wu-Scharf *et al.*, 2000). This is thought to be a defence mechanism against lysogenic viral attack, which explains why this process is not observed for the chloroplast, which no viruses are known to target. Another key factor in recombinant protein accumulation is the chloroplast protease environment. The chloroplast contains a suite of prokaryotic-type proteases that might be less active on foreign proteins than those encountered in the cytoplasm (Bock, 2001). The chloroplast moreover allows for direct targeting of gene insertion into the plastome via homologous recombination, eliminating many of the difficulties associated with positional effects of randomly inserted transgenes (Purton, 2007). The major features of both nuclear and chloroplast transgene expression are summarised in Table 1.4.

Owing to the endosymbiotic nature of the chloroplast, it is in a unique position between the pro- and eu-karyotic worlds: it has retained much of its prokaryotic expression machinery such as the 70S ribosome and as such shows good expression of bacterial genes. It also has various eukaryotic characteristics, such as the capacity to correctly assemble complex eukaryotic structures incorporating di-sulphide bridges, such as antibodies (Mayfield *et al.*, 2003). This allows for interesting, slightly paradoxical applications, such as the concept of using the chloroplast to synthesise antibody-linked eukaryotic toxins for targeting cancerous cells. Conventional systems would abhor such a fusion, with bacterial platforms being unable to fold the antibody, and eukaryotic ones unable to tolerate the toxin; however, the algal chloroplast is fully capable of such synthesis (Tran *et al.*, 2013). Furthermore, as the chloroplast is primarily an anabolic organelle and is partitioned from the activity of the cytoplasm, it can be seen as relatively

metabolically 'safe' allowing greater accumulation of recombinant product than might otherwise be possible (Mayfield *et al.*, 2007).

Table 1.4 - A comparison of nuclear and chloroplast expression of transgenes

	Nuclear Expression	**Chloroplast Expression**
Mode of insertion	Random	Targeted (homologous recombination)
Copy number of transgene	One + (although multiple insertions can activate silencing mechanisms)	One per genome; ~80-100 per cell
Gene silencing	Transcriptional and post-transcriptional	None observed
Positional effects	Random insertions result in variable expression levels	No positional effects due to targeted insertion to neutral loci
DNA packaging	Extensive chromatin structures	Basic packaging around bacteria-like nucleoids
Glycosylation	Similar to higher plants	None observed
Location of primary product	Cytosol	Chloroplast
Potential for export	Can be targeted to the chloroplast, ER, or for secretion	None
Maximum reported expression	0.25 % TSP	10.5 % TSP

1.3.5 C. reinhardtii as a platform for endolysin expression

It is precisely this dichotomy that makes the chloroplast of *C. reinhardtii* ideal for lysin production. In many cases, lysin expression begins well in advance of cell lysis and hence these enzymes are exposed directly to the host cytoplasm for much of the phage lifecycle (Fischetti *et al.*, 2006; Wang *et al.*, 2000). Owing to the highly evolved nature of the bacteriophages as organisms, this has resulted in their lysins displaying a remarkable resistance to bacterial proteases, and hence a high degree of stability in the bacterial cytoplasm. It is now widely accepted that the evolution of the green algal chloroplast is the result of a primary endosymbiotic event approximately 1.5 billion years ago as illustrated in Figure 1.9 (Gray and Doolittle, 1982; Keeling, 2004). The proteases seen in the chloroplast can thus be thought of as a result of the alga's prokaryotic ancestry, as has been demonstrated by homology studies (Adam *et al.*, 2006). It is consequently assumed that the high protease resistance shown by lysins in their native hosts will be reflected when synthesised in the chloroplast. This was recently demonstrated by Oey and colleagues: the *Streptococcus pneumoniae* specific lysin PlyGBS was expressed in the tobacco chloroplast and was found to accumulate to 70 % of total soluble protein. These extreme expression levels were shown to be largely due to little or no turnover of the protein (Oey *et al.*, 2009a). The use of a chloroplast based system allows for this prokaryote tailored stability to be exploited without the difficulties presented by bacterial endotoxins, or cross reactivity of the lysin with its natural substrate, peptidoglycan, which is not found in the chloroplast of green algae or higher plants (although it is seen in the chloroplast of glaucophytes). The presence of peptidoglycan *per se* in bacterial platforms should not be of consequence due to the specific nature of the lysins; however, host toxicity in bacteria has been reported at high expression levels (Oey *et al.*, 2009b). There is also considerable interest in using synthetic biology to expand lysin target ranges, which may in turn lead to host peptidoglycan reactivity in a bacterial expression platform. A platform where the target substrate is completely absent can be considered to be 'future-proof' in this respect.

The combination of factors discussed herein therefore makes the *C. reinhardtii* chloroplast an ideal expression platform for the production of the novel class of antibiotics, the endolysins.

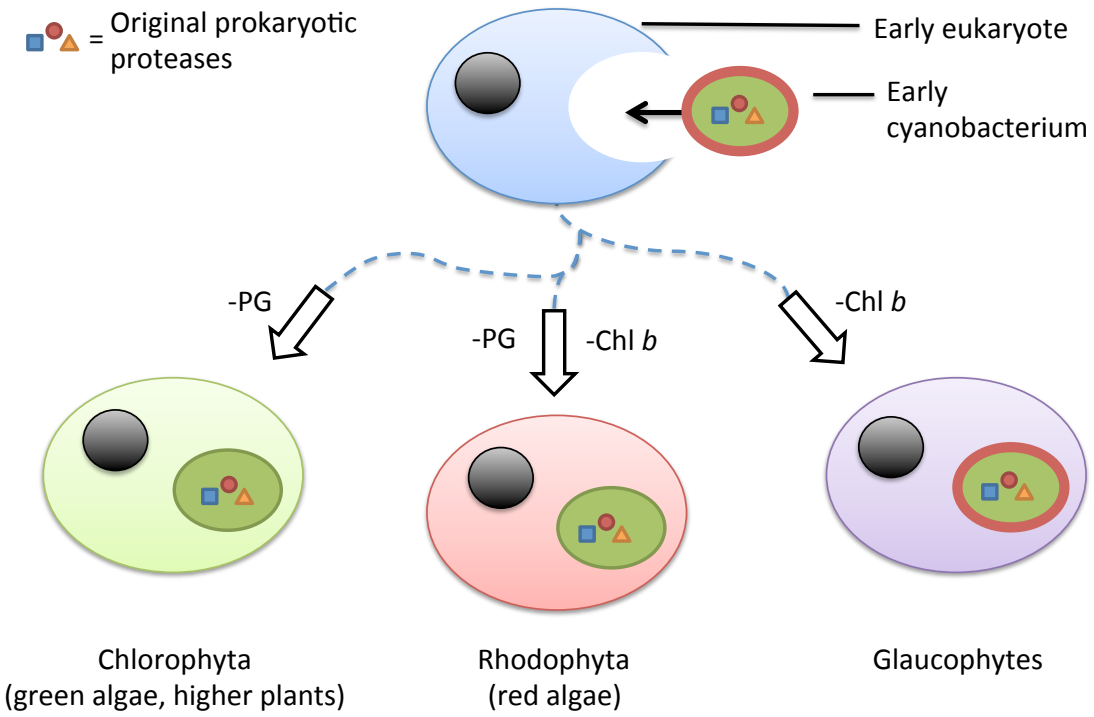


Figure 1.9 – Simplified schematic showing the primary endosymbiotic event thought to have given rise to the first plastids

The proteases in each resulting plastid can be seen to be descendent from the original cyanobacterial symbiont and thus show closer identity to prokaryotic rather than eukaryotic homologues. Abbreviations: PG - peptidoglycan cell wall, Chl - chlorophyll.

1.4 Improving the expression of transgenes in the *C. reinhardtii* chloroplast

1.4.1 Limitations to commercial uptake: low product yields and poor reliability of transgene expression

As discussed above, *C. reinhardtii* represents a highly promising platform for the production of recombinant proteins, not least due to its GRAS status, ability to fold complex proteins, and capacity to store product in the relatively metabolically safe chloroplast. Despite these and other advantages, *C. reinhardtii* has yet to become established as a routine expression platform for therapeutic (or indeed other commercially interesting) proteins. A major contributing factor is one of protein productivity; transgenic *C. reinhardtii* does not produce enough recombinant protein relative to mature platforms for it to be an economically viable option. Combined with the added costs inherent in pioneering a non-standard technological platform, this currently represents a major economic barrier to commercial adoption of *C. reinhardtii* for recombinant protein synthesis. Considerable effort has been invested into improving the expression of transgenes in *C. reinhardtii*, with research focusing on the chloroplast due to its intrinsically higher yields.

1.4.2 Approaches to improving transgene expression in *C. reinhardtii* relative to the central dogma of gene expression

The transformation of *C. reinhardtii* was first achieved in the late 1980s for both the chloroplast (Boynton *et al.*, 1988) and nuclear genomes (Kindle *et al.*, 1989). However, numerous potential pitfalls exist between the integration of foreign DNA and the successful accumulation of recombinant protein; Figure 1.10 summarizes the standard dogma required for successful gene expression as well as a selection of 'molecular snags' that must be avoided. The major stages in the gene expression pathway, and current progress in their optimization, are discussed below.

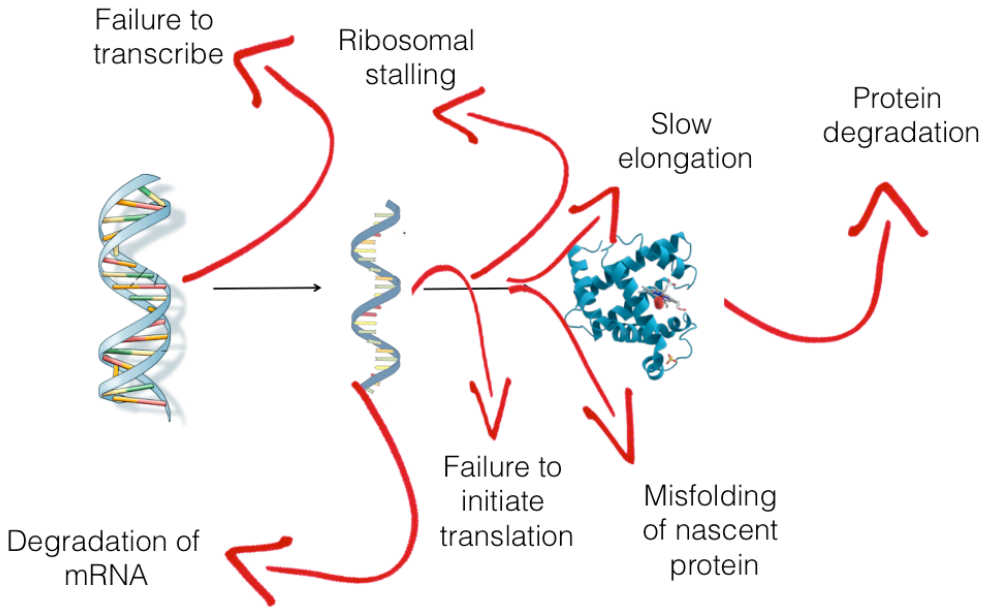


Figure 1.10 – Potential 'molecular snags' in the transition from DNA to RNA to protein.

Once a transgene has been stably transformed into an organism there are still a number of issues to be overcome before recombinant protein can be detected, a selection of which are illustrated.

1.4.3 Promoter selection

A common approach to heterologous protein expression is to use highly expressing viral promoters such as the T7 virus promoter in bacterial systems and the cauliflower mosaic virus (CMV) 35S promoter in many higher plants. To date no heterologous promoter has been shown to be active in the *C. reinhardtii* chloroplast (Purton, 2007; Walker *et al.*, 2005). Endogenous promoters however have been shown to be effective, with the promoters from the highly expressed photosynthetic genes *psbA, atpA*, and *psaA* all giving successful expression of transgenes in the chloroplast (Purton *et al.*, 2013; Rasala *et al.*, 2011). Furthermore it has been demonstrated that endogenous gene expression in the *C. reinhardtii* chloroplast is primarily regulated at the post-transcriptional level (Eberhard *et al.*, 2002; Hosler *et al.*, 1989) and transgenic transcripts have been shown to be in excess (Coragliotti *et al.*, 2011), discouraging further promoter optimisation for increased yield. Though functional, use of endogenous promoters is not ideal. Due to the prolific homologous recombination that occurs within the chloroplast genome, multiple copies of a promoter are likely to induce genomic instability (Stern and Harris, 2009). There is also a shortage of effective inducible systems for controlled induction of transcription in the algal chloroplast (Purton *et al.*, 2013).

1.4.4 5' and 3' untranslated region selection and optimisation

5' and 3' untranslated regions (UTRs) have two main functions: to stabilise the mRNA transcript, and, in the case of the 5' UTR, to initiate translation. For transgene expression in the *C. reinhardtii* chloroplast it has been shown that variation of the 3' UTR has little effect on recombinant protein accumulation, with the proviso that there is a 3' UTR present (Barnes *et al.*, 2005). 5' UTR optimisation has proven to be considerably more complex.

In general, prokaryotic translation initiation is dependent on successful assembly of the 70S ribosome translation initiation complex around the translation initiation region (TIR) of the 5' UTR followed by the correct recognition of the start codon. These two occurrences are governed by four factors: the Shine Dalgarno (SD) sequence, the start codon, the spacing between the SD and the start codon, and any up or downstream enhancer elements present (Huang *et al.*, 2012). Chloroplast

translation is essentially prokaryotic in nature, as has been exploited in higher plant systems where bacterial 5' UTRs such as gene 10 leader from the T7 phage have driven strong transgene expression (Oey *et al.*, 2009a). Such systems have not been reported to be active in the *C. reinhardtii* chloroplast however, and it is thought this is in a large part due to the tight control over chloroplast gene expression emanating from the nucleus (Specht *et al.*, 2010). Recently, such mechanisms have started to be elucidated. Control of translation initiation can broadly be split into two functional moieties: direct secondary mRNA structures which interact with the initiation complex, such as the hairpin repressors seen in the *psbD* 5' UTR (Klinkert *et al.*, 2006), and elements which recruit nuclear-encoded *trans* acting factors, such as the RB47 translational activator binding site of *psbA* (Yohn *et al.*, 1998). Such *trans* acting factors can then act in two ways: by stabilising the mRNA, and/or directly aiding (or blocking) ribosome recruitment.

The two reports of highest accumulation of recombinant protein in the *C. reinhardtii* chloroplast were both based on expression using the *psbA* promoter/ 5' UTR combination (Manuell *et al.*, 2007; Surzycki *et al.*, 2009). This is unsurprising as *psbA* has been shown to be the most highly expressing endogenous chloroplast gene; its product, D1, is rapidly turned over due to frequent photo-oxidative damage (Weiß *et al.*, 2012). Interestingly, these high levels of recombinant protein accumulation are only seen in *psbA* deficient, and thus non-photosynthetic, strains; however, when *psbA* is re-introduced under the less active *psbD* promoter/ 5' UTR much of the transgene productivity seen in *psbA*-deficient lines is retained (Manuell *et al.*, 2007). It was subsequently shown that D1 directly auto-regulates *psbA* expression (and by extension down-regulates any transgenes driven by the *psbA* elements) via interaction with the *psbA* 5' UTR (Rasala *et al.*, 2011). The process of D1/*psbA* auto-attenuation is summarised in Figure 1.11.

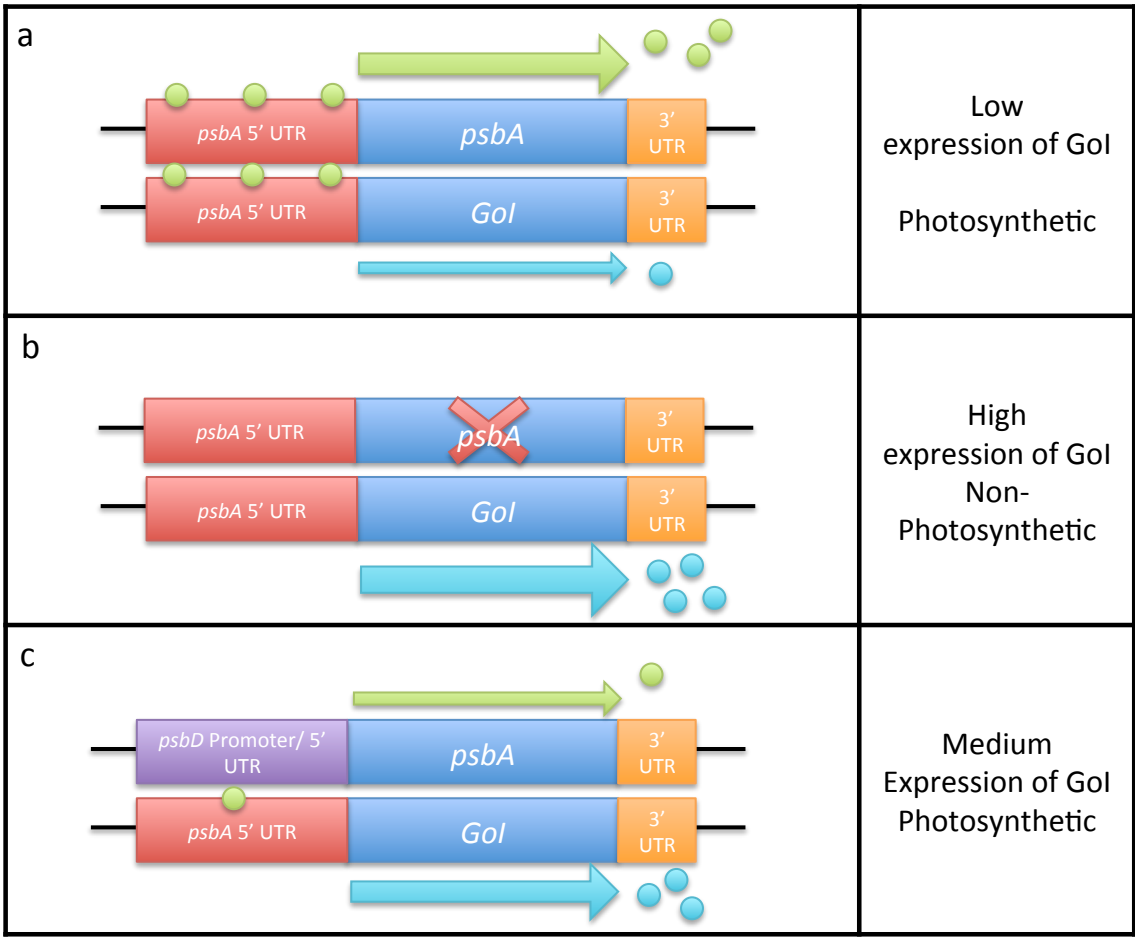


Figure 1.11 – A simplified summary of *psbA*/D1 auto-attenuation in regard to transgene expression

In panel **(a)** the *psbA* gene product D1 (green balls) regulates both *psbA* and GoI expression. Recombinant product (cyan balls) accumulation is low. In panel **(b)** *psbA* has been knocked out and photosynthetic activity removed. D1 mediated down-regulation of the GoI is lifted and high expression observed. In panel **(c)** the *psbA* gene is re-introduced under the control of the less active *psbD* promoter/ 5' UTR and photosynthesis is restored. Transgene expression is reduced relative to **(b)** but is still higher than **(a)**.

Until recently such investigations into enhancer/ repressor 5' UTR elements were conducted on a qualitative basis with little comment made as to the mechanism involved or how 5' UTRs could be rationally designed to avoid such attenuation. Research published this year by Specht and colleagues has challenged this paradigm (Specht and Mayfield, 2013). In a highly ambitious work over 900 variants of the *psbD* and *psaA* 5' UTRs were cloned into the *C. reinhardtii* chloroplast and analysed using a luciferase reporter. These mutants were then analysed for reporter content and split into two groups of high and medium/low expressers. From base-by-base analysis of percentage wild type conservation in the two groups the authors were able not only to precisely identify the key *cis* acting elements in the two 5' UTRs, but also to verify their findings by generating a synthetic 5' UTR which went on to give strong expression of a recombinant gene. A further finding was that while they were unable to find conservation of a typical Shine-Dalgarno motif in the -10 region, SD-like sequences were found to be heavily conserved in high expressing groups at -29 and -51 for *psbD* and *psaA,* respectively. Such SD position flexibility has previously been reported in other chloroplast genomes (Harris *et al.*, 1994).

An area not included in this investigation was the effect of sequences downstream of the AUG translation start. Such regions can be of importance for two reasons. Firstly, in terms of direct interaction with local sequence, for example the second hairpin loop seen in *psbD,* which forms between -4 and +12 (Klinkert *et al.*, 2006). Secondly, in regard to a region referred to as the downstream box (DB), an area of ribosome-interacting bases ranging from +15 to +26 (Sprengart *et al.*, 1996). In *E. coli* the DB has been shown to interact with a complementary region of the 16S ribosomal rRNA known as the anti-downstream box, ranging from 1469-1483, and has been shown to actively regulate translation efficiency. Isolated cases have shown inclusion of a downstream box from highly expressing genes to increase recombinant chloroplast expression (Gray *et al.*, 2011; Kasai *et al.*, 2003), however this introduces complications relating to N-terminal extensions of the protein of interest.

1.4.5 Translational elongation and codon optimization

Translation elongation relates to the speed at which the nascent polypeptide is formed and is key to the regulation of protein expression in two ways: firstly by controlling how fast a protein is produced (in non initiation limited systems), and secondly by subtly controlling the rate of protein synthesis to assist with correct folding of the peptide. This can be achieved, for example, by reducing speed and/ or pausing elongation to allow the folding of one domain before continuing to the next (Welch *et al.*, 2009b). Rates of translation elongation are widely thought to be related to the degree of codon optimisation of the gene relative to the host organism (Plotkin and Kudla, 2011). The basic premise of codon optimisation is that the redundancy in the genetic code allows for multiple codons to encode the same amino acid, but different organisms have their own preferences as to which are actually used. These preferences are based predominantly on three factors: the relative concentrations of the corresponding tRNAs available, the recharge rate of tRNAs, and the stability of the transient ribosome:tRNA complex. By matching a transgene's codon usage to that of the expression platform it is thought that optimal translation optimisation can be achieved (Plotkin and Kudla, 2011; Potvin and Zhang, 2010).

In the wider academic community there is a degree of controversy over the importance of codon optimisation in recombinant gene expression (Gustafsson *et al.*, 2012; Kudla *et al.*, 2009; Welch *et al.*, 2009a). Reports involving codon optimisation of transgenes for *C. reinhardtii* expression, however, have consistently indicated positive results. Early work on transgene expression using the *gfp* gene encoding green fluorescent protein as a reporter showed significant increases in product accumulation for both the nucleus (5 fold increase, (Fuhrmann *et al.*, 1999)) and the chloroplast (80 fold increase, (Franklin *et al.*, 2002)). Several other studies have shown positive results for optimised transgenes including human antibodies and luciferase (Mayfield and Schultz, 2004; Mayfield *et al.*, 2003). A recent study by Coragliotti and colleagues has also demonstrated that in the case of three transgenes investigated, translational elongation was the major limiting factor in expression (Coragliotti *et al.*, 2011), reinforcing the importance of correct codon usage.

As well as 'good' codons being shown to increase expression, it has been observed that rare codons can prevent the expression of proteins in the *C. reinhardtii* chloroplast. The effect of rare codons was investigated by Weiβ and colleagues by creating specific rare codon mutations in a flexible loop region of D1 where it had previously been shown that amino acid alterations were tolerated. These constructs were then used to restore non-photosynthetic mutants lacking a functional *psbA* gene and thus D1 protein. It was found that although many rare codon combinations were tolerated, combinations of the rare arginine codons CGG and AGG were not, as seen by a complete inability to isolate transformants containing these constructs. Certain combinations of rare codons such as triplets of the rare serine codon TCC were also shown to be able to cause ribosome pausing (Weiß *et al.*, 2012).

An area of translation elongation yet to be investigated for *C. reinhardtii* is that of codon pair optimisation. It has long been known that in *E. coli* the distribution of codon pairs is not as random as would be expected if only individual codon use had a bias acting upon it (Gutman and Hatfield, 1989). Studies investigating the finer points of codon pairing have shown that the ribosome decodes preferred codon pairs faster than 'bad' ones (Boycheva *et al.*, 2003) with the rationale that the simultaneous occupation of the A and P site in the ribosome dictates that the stability of the complex should be determined by both occupying tRNAs in concert with the ribosome. Although codon usage is an established concept in *C. reinhardtii* recombinant gene expression, to the author's knowledge codon pair effects have yet to be investigated.

1.4.6 Additional factors in transgene expression

In addition to the factors discussed above, several other processes are thought to influence transgene expression in the *C. reinhardtii* chloroplast. Active areas of investigation include chloroplast protease function (Adam *et al.*, 2006) and assisted folding by recombinant chaperones (C. Economou, unpublished work). Such studies are hindered however by a lack of knowledge regarding the basic biology behind these processes in the *C. reinhardtii* chloroplast.

1.5 Summary, Aims, and Objectives

In the course of reviewing the relevant literature, it has been shown that there is an urgent need for a novel class of antibiotics that addresses the issues of resistance development, damage to the natural microbial flora, and ability to target mucosal membranes and biofilms. The advantages of protein-based therapeutics have been discussed, and the bacteriophage endolysins proposed as a possible solution to many of the problems identified. The demand for new, low cost expression platforms for biologic production has been explored, with the use of plants as protein factories revealed to show potential; however, the green alga *C. reinhardtii* is presented as the most suitable platform on grounds of containment, rate of growth, and minimal requirements for cultivation. Issues of low yields have been considered, and current progress in improving productivity reviewed.

To address the issues thus raised the following aims formed the basis of the research described in this thesis:

1. To express one or more bacteriophage lysins in the *C. reinhardtii* chloroplast, demonstrating antimicrobial activity and developing methods for their purification

2. To investigate factors affecting expression of recombinant genes in the *C. reinhardtii* chloroplast, focusing on the re-creation of entire translation initiation regions from highly expressing genes

3. To explore the nature of codon- and codon pair bias in the *C. reinhardtii* chloroplast, relating the findings to the design of synthetic transgenes

Chapter 2

Methods

During the course of the projects discussed in this thesis some protocols were optimised with parameters varied. The methods herein described are the optimised procedures and thus may not be identical to those used in the following results chapters. Where possible such differences are noted in the relevant results sections.

2.1 Strains, culture conditions, and quantification of cell density

2.1.1 Chlamydomonas reinhardtii

Wild type *Chlamydomonas reinhardtii* CC-1021 (mt+) was obtained from the Chlamydomonas Resource Center (www.chlamy.org) and is a backcross of two wild-type strains that originate from the founder isolate 137C (Spreitzer and Mets, 1981). The recipient line bst-same1 was created by O'Connor *et al*, by insertional knockout of *psbH* from strain CC-1021 (O'Connor *et al.*, 1998). The improved recipient line TN72 was developed in the Purton lab by Dr Thanyanun Ninlayarn by insertional knockout of *psbH* from the cell wall-less strain cw15(mt+) (unpublished work).

C. reinhardtii strains were cultured in either Tris-Acetate Phosphate (TAP) or High Salt Minimal (HSM) growth medium depending on the application (Table 2.1). Liquid cultures were grown under 24 hour illumination at 25 °C with 125 rpm rotary agitation and an average light intensity of 25–50 μmol/m^2/s in an illuminated incubator shaker (Innova 4340, New Brunswick Scientific). Starter cultures were made by inoculating 25 ml TAP medium with a large loopful of actively growing *C. reinhardtii* and growing to mid-log phase (approximately 5 x 10^6 cells) at which point the culture was sub-cultured at a 100-fold dilution.

Growth on nutrient agar plates was on TAP medium supplemented with 2 % bacto agar (w/v) at 25 °C and an average light intensity of 50 μmol/m^2/s. Long term storage of strains was conducted on TAP medium supplemented with 2 % bacto-agar in dim light (5 μE/m^2/s) at 18 °C with re-streaking to fresh TAP plates every 6–8 weeks.

Cell density was determined using a haemocytometer, counted by eye at 400 X magnification. Motile cells were first treated with 10 µl tincture of iodine (19.7 mM iodine in 95 % (v/v) ethanol) per 1 ml sample to facilitate immobilisation. Optical density measurements were made at 750 nm (to avoid the majority of chlorophyll *a* and *b* absorbance), with a path length of 1 cm. The spectrophotometer used in all cases was a Unicam UV/Vis Spectrometer UV2.

Table 2.1 – Tris-acetate phosphate (TAP) and high salt minimal (HSM) growth media recipes for *C. reinhardtii*. Adapted from (Rochaix *et al.*, 1988)

[a]4 x Beijerinck salts: 0.3 M NH_4Cl, 14 mM $CaCl_2{\cdot}2H_2O$ and 16 mM $MgSO_4{\cdot}7H_2O$

[b]1 M (K)PO_4, pH 7: 1 M K_2HPO_4 titrated to pH 7.0 with 1M KH_2PO_4

[c]2 x PO_4 for HSM: 80 mM K_2HPO_4 and 50 mM $KH_2\ PO_4$, adjusted to pH 6.9 with KOH

[d]Trace elements: 180 Mm H_3BO_3, 77 mM $ZnSO_4{\cdot}7H_2O$, 26 mM $MnCl_2{\cdot}4H_2O$, 18 mM $FeSO_4{\cdot}7H_2O$, 7 mM $CoCl_2{\cdot}6\ H_2O$, 6 mM $CuSO_4{\cdot}5H_2O$, 0.9 mM $(NH_4)_6Mo_7O_{24}{\cdot}4H_2O$.

For 1 litre	Tris-acetate phosphate (TAP)	High salt minimal (HSM)
H_2O	975 ml	925 ml
Tris	2.42 g	-
[a]4 x Beijerinck salts	25 ml	25 ml
[b]1 M (K)PO_4, pH 7	1 ml	-
[c]2 x PO_4 for HSM	-	50 ml
[d]Trace elements	1 ml	1 ml
Glacial acetic acid	to pH 7.0 (approximately 1 ml)	-

2.1.2 Escherichia coli

All bacterial DNA amplification and prokaryotic protein expression was conducted using *E. coli* strain DH5α, genotype F-, (φ80*lacZΔM15*), Δ(*lacZYAargF*) U169, *deoR*, *recA1*, *endA1*, *hsdR17*(rk-, mk+), *supE44*, *thi-1*, *gyrA96*, *relA*, λ (Hanahan, 1983). The strain was supplied by Clontech (Saint-Germain-en-Laye, France). *E. coli* was grown in Luria-Bertani (LB) medium (Table 2.3) supplemented with 100 μg/ml ampicillin or kanamycin when necessary. Liquid cultures were grown overnight at 37 °C with 200 rpm rotary agitation in a shaking incubator (Innova 4430, New Brunswick Scientific). Inoculation was from a single colony using a sterile toothpick. Colonies were grown on LB medium supplemented with 2 % Difco agar at 37 °C overnight. *E. coli* strains were stored as 20 % (v/v) glycerol stocks at -80 °C. For all bacterial cultures, cell density was measured by optical density at 600 nm.

2.1.3 Streptococcus pneumoniae

S. pneumoniae strains were acquired from the Royal Free hospital and are listed in Table 2.2. *S. pneumoniae* was cultured overnight at 35 °C without shaking in a 6 % CO_2 enriched atmosphere generated using Oxoid AGS CO_2Gen Compact gas packs. Liquid cultures were grown in Trypticase soy yeast extract medium (TSY) medium (Table 2.3). Inoculation was from a single colony using a sterile toothpick. Colonies were grown on Columbia Blood Agar (CBA) supplied premade from Sigma under the same conditions as described for liquid media growth. *S. pneumoniae* strains were stored in skimmed milk tryptone glycerol glucose (STGG) medium (Table 2.3) at -80 °C.

2.1.4 Staphylococcus aureus

S. aureus strain ATCC 28213 was cultured overnight at 35 °C without shaking. Liquid cultures were grown in commercially purchased Oxoid Iso-SensiTest broth medium (Table 2.3). Inoculation was from a single colony using a sterile toothpick. Colonies were grown on Columbia Blood Agar (CBA) supplied premade from Sigma. *S. aureus* strains were stored in Oxoid iso-sensitest broth medium supplemented with 20 % (v/v) glycerol at -80 °C.

Table 2.2 – Strains of *S. pneumoniae*

All strains were acquired from the Clinical Microbiology department, Royal Free Hospital

Serotype	Strain	Identifier
6A	ST65	H08212 0259
6B	ST176	H08052 0052
6C	ST1390	H05252 0075
19A	16NP3	--
27	ST1475	H08432 0293

Table 2.3 – Growth media recipes for bacterial cell culture.

Organism	Medium	Recipe	Reference
E. coli	Luria-Bertani (LB)	1 % (w/v) bactotryptone, 0.5 % (w/v) bacto-yeast extract, 0.17 M NaCl	(Bertani, 1951)
S. pneumoniae	Trypticase soy yeast extract (TSY)	3 % (w/v) Trypticase soy broth, 0.3 % (w/v) Yeast extract pH 7.0	http://www.dsmz.de/?id=441
S. pneumoniae	Skimmed milk tryptone glycerol glucose (STGG)	2 % (w/v) skimmed milk powder, 3 % (w/v) Tryptone soy broth, 0.5 % (w/v) glucose 10 % (v/v) glycerol	(O'Brien *et al.*, 2001)
S. aureus	Oxoid iso-sensitest broth	Commercially purchased (Thermo Scientific)	(Snyder and Atlas, 2006)

2.2 Molecular biology

2.2.1 Gene design, optimisation, and synthesis

All gene sequences were acquired from Genbank[1] and verified by Blastn search[2] prior to optimisation and synthesis. Codon optimisation was conducted either by the gene synthesis company, GeneArt[3] or in-house using the Codon Usage Optimizer software[4].

GeneArt optimisation was conducted to a target Codon Adaptation Index (CAI) of 0.8 using the *C. reinhardtii* chloroplast codon usage table provided by the Kazusa institute (Nakamura *et al.*, 2000). Sequences were optimised by GeneArt using the online optimiser associated with their gene synthesis service. CUO optimisation was conducted using the Codon Usage Optimizer software, developed by Khai Kong in the Purton lab. A target CUO of 1 was used; however, as this method also takes codon pairing into account, a compromise is met between ideal codon use and non-deleterious codon pair use. Codon- and codon pair tables were generated using a selection of highly expressed *C. reinhardtii* chloroplast genes, a subset referred to as *C. reinhardtii chloroplast handpick.*

All genes were synthesised by GeneArt and received as plasmid clones with the gene inserted into either of the company's pMA or pMK vectors that carry a selectable marker for ampicillin or kanamycin resistance respectively.

2.2.2 Isolation of *C. reinhardtii* genomic DNA

Crude genomic preparations for PCR verification of chloroplast transformants were conducted using a rapid isolation protocol. A small loopful of actively growing cells was resuspended in 10 µl ddH_2O to which 10 µl absolute ethanol was added and the mixture incubated at room temperature for 1 minute. 200 µl of a 5 % (w/v) Chelex-100 resin (Biorad) suspension was then added and the sample

[1] http://www.ncbi.nlm.nih.gov/genbank

[2] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[3] http://www.invitrogen.com/geneart

[4] http://codonusageoptimizer.org

vortexed at full speed for 15 seconds. Samples were then placed in a heat block (Eppendorf Thermomixer comfort 1.5 ml) at 98 °C for 5 minutes before cooling on ice and centrifuging (2 minutes, 21k x *g*). 2 µl of supernatant was used per PCR reaction. Remaining supernatant was discarded or split into aliquots and stored at -20 °C for a maximum of 2 months before being discarded. Thawed samples were discarded after use.

More stable long-term preparations for repeated use as positive and negative controls were made using a 'mini-prep' protocol adapted from Rochaix *et al.* (Rochaix *et al.*, 1988). *C. reinhardtii* cells were grown to a density of approximately 3 x 10^6 cells/ml and harvested by centrifugation at 4k x *g* for 5 min. The pellet was resuspended in 0.35 ml of resuspension solution (50 mM EDTA, 20 mM Tris-HCl, pH 8.0, 0.1 M NaCl), and transferred to a 1.8 ml eppendorf tube. Proteinase K and SDS were then added to a final concentration of 0.3 mg/ml and 1.4 % (w/v), respectively, and the sample incubated at 55 °C for 2 hours. Diethylpyrocarbonate (DEPC, 2 µl) was then added and the mixture incubated at 70 °C for 15 minutes. The sample was cooled on ice and 50 µl 5 M potassium acetate added prior to incubation on ice for a further 30 minutes. Debris was removed by centrifugation (15 minutes, 21k x *g*), and the supernatant mixed with an equal volume of phenol in a fresh 1.8 ml eppendorf tube. The sample was vortexed for 1 minute, centrifuged (2 minutes, 21k x *g*), and the upper aqueous phase retained. DNA was then precipitated by adding 1.5 ml absolute ethanol, centrifuged (2 minutes, 21k x *g*), and washed with 1.5 ml 70 % (v/v) ethanol. After a final centrifugation (1 minute, 21k x *g*) the pellet was dried and resuspended in 50 µl TE buffer (10 mM Tris-HCl, pH 7.5, 1 µg/ml pancreatic RNase) and stored at -20 °C.

2.2.3 *E. coli* genomic preparations for colony PCR

Confirmation of *E. coli* transformation was initially by colony PCR (cPCR). Colonies were picked using a sterile toothpick and resuspended in 20 µl of a pre-prepared PCR reaction mix. Once resuspended, the toothpick was then transferred to 5 ml LB amp^{100} in a 20 ml Sterilin tube, and incubated for 6 hours to allow growth from the same colony. PCR was conducted as described below with the addition of an initial cell lysis stage of 2 minutes at 98 °C. Confirmed transformants were sub-cultured from the 6 hour incubations.

2.2.4 Isolation of plasmids from *E. coli*

Small scale (<12 μg) isolation of plasmids was conducted using the Qiagen Miniprep kit as per the manufacturer's instructions. In brief, 1.5 ml of overnight *E. coli* culture was broken by alkali lysis, genomic DNA and protein precipitated by potassium acetate, and the supernatant loaded onto a silica based affinity spin column. After washing with an ethanol based wash solution, the plasmid DNA was eluted in 30 μl TE buffer.

Larger scale isolations (<200 μg) were conducted using the Qiagen Midiprep kit as per the manufacturer's instructions. In brief, 25 ml of *E. coli* overnight culture was broken as for the miniprep protocol. Precipitated genomic DNA and protein was then removed by filtration, and the flow through loaded onto an ion exchange column. DNA was eluted with 5 ml 1.25 M NaCl elution buffer and the elutant concentrated by ethanol precipitation.

2.2.5 Polymerase chain reaction

DNA amplification for both diagnostic and cloning purposes was conducted using the polymerase chain reaction (PCR). The primers used for each reaction are listed in appendices for the appropriate Results chapter. Standard reactions were 50 μl in volume and comprised: 0.5 μl 20 mM mixed dNTPs (final concentration 200 μM dATP, 200 μM dTTP, 200 μM dGTP and 200 μM dCTP), 0.5 μl forward and reverse oligonucleotide primers (100 pmol/μl), 2 μl template DNA (approximately 1 μg of genomic DNA or 20 ng of plasmid DNA), 10 μl 5 x Phire II reaction buffer (New England Biolabs), 0.5 μl Phire II Hot-Start DNA polymerase (1 U) and ddH_2O to make the final volume 50 μl. Smaller reactions of 20 μl were scaled down accordingly. The DNA was amplified using a Techne TC-3000X thermocycler. The reactions were denatured for 30 s at 98 °C and subjected to 25 cycles of 5 s denaturation at 98 °C, 5 s of annealing at approximately 3 °C below the average T_m of the two primers, and extension at 72 °C for 15 s/ 1kb of amplicon. The reactions were subjected to a final incubation at 72 °C for 2 minutes and held at 4 °C. Analysis of the PCR product obtained was by agarose gel electrophoresis of 10 μl of the PCR reaction.

2.2.6 Restriction endonuclease digestion

DNA samples were digested using 10 units of restriction endonuclease (New England Biolabs) per μg of DNA according to the manufacturer's instructions. In cases of double digests the suitable buffer and ratio of each enzyme were determined using the NEB Double Digest Finder[5]. The exceptions to this were *Sap*I and *Sph*I, which were supplied by Fermentas due to improved double digest compatibility.

2.2.7 Agarose gel electrophoresis

DNA fragments were separated on 1 % (w/v) agarose gels supplemented with 0.1 μg/ml ethidium bromide. 6x loading dye (2.5 % (w/v) Ficoll 400, 11 mM EDTA, 3.3 mM Tris-HCl, 0.017 % (w/v) SDS and 0.015 % (w/v) Bromophenol Blue) was added to DNA samples to a 1x final concentration. Commercial size markers (New England Biolabs or Fermentas) were run on all gels at 0.5 μg per lane. Gels were made to a total volume of 50 ml and were submerged in TAE buffer (0.04 M Tris, 1 mM sodium EDTA and 17.5 mM glacial acetic acid) in electrophoresis tanks supplied by Sigma. The gels were run at 100 V, supplied by a Gibco BRL power supply. Gels were visualized on a UV illuminator and images captured on black/white thermal paper (UVP Gel Documentation System).

2.2.8 Removal of 5' phosphate from DNA using Antarctic phosphatase

Antarctic phosphatase was used to treat restriction-digested plasmid to prevent re-forming of the parent plasmid in subsequent ligation reactions. DNA samples were treated using 5 units of Antarctic phosphatase (New England Biolabs) per 1 μg of DNA according to the manufacturer's instructions. In brief, the reaction was incubated for 1 hour at 37 °C and heat deactivated for 5 minutes at 65 °C.

2.2.9 Purification of PCR product

PCR purification was employed prior to sequencing to remove primers, buffer and polymerase from PCR reactions. In addition, the same method was used to remove small (<50 bp) DNA fragments released from plasmids following restriction digest.

[5] https://www.neb.com/tools-and-resources/interactive-tools/double-digest-finder

This PCR 'clean-up' was conducted using the Qiagen PCR clean-up kit as per the manufacturer's instructions. In brief, the sample was added to an equal volume of binding buffer and loaded onto a silica affinity spin column. Washing and elution of DNA was conducted as for the Qiagen Mini-prep kit above.

2.2.10 DNA Ligation

For ligation reactions, a molar insert:vector DNA ratio of approximately 3:1 was used. Ligation was conducted with 1 U of T4 DNA ligase (New England Biolabs) per 1 μg DNA for 1 hour at room temperature. Total reaction volume was 10 μl.

2.2.11 DNA Sequencing

DNA was sequenced by an in-house UCL service based at the Wolfson Institute for Biomedical Research using a Beckman Coulter CEQ 8000 genetic analysis system. Primers were supplied for sequencing at a concentration of 2-5 pmoles/μl. Plasmid and PCR product DNA were supplied at a concentration of 9-16 fmoles/μl and 4-8 fmoles/μl, respectively.

2.3 Genetic transformation

2.3.1 Escherichia coli

2.3.1.1 Preparation of competent E. coli cells

E. coli (DH5α) cells were restreaked on LB plates from a master glycerol stock stored at -80 °C, and cultured overnight at 37 °C. 10 ml liquid cultures were inoculated from a single colony and grown overnight as described above (section 2.1.2). From this a 100 ml liquid culture was inoculated at a 100-fold dilution and grown for 2.5 hours at 37 °C to an optical density of approximately 0.6 at 600 nm. The culture was then cooled on ice and the cells centrifuged at 4k x *g* for 5 minutes in four pre-cooled sterilin tubes. The pellets were each resuspended in 10 ml ice cold 50 mM $CaCl_2$, incubated on ice for 30 minutes, and centrifuged at 4k x *g* for 5 minutes. Pellets were resuspended in 1.5 ml of fresh 50 mM $CaCl_2$. Samples were pooled and 3.5 ml of sterile 50 % (v/v) glycerol added. The cells were dispensed into microcentrifuge tubes in 250 μl aliquots and frozen at -80 °C until required.

2.3.1.2 ***Transformation of* E. coli**

Aliquots of competent cells prepared as described above were thawed on ice and split into sterile 1.8 ml eppendorf tubes with 100 µl used per transformation. For transformation with ligation mixes, the entire 10 µl reaction was added to the competent cells, whereas for transformations with intact plasmid a miniprep of the plasmid DNA was diluted by a factor of 10 and 1 µl used. A negative control with no addition of DNA to the tube was always included. Samples were incubated on ice for 30 min then heat shocked at 42 °C for 60 s and immediately returned to ice. To each tube, 1 ml of LB medium was added prior to incubation at 37 °C for 1 hour. From each sample, 200 µl was spread onto an LB amp^{100} agar plate and incubated overnight at 37 °C. For transformations involving ligations, the remaining 900 µl was centrifuged, resuspended in 200 µl of supernatant, and plated.

2.3.2 **The chloroplast of *Chlamydomonas reinhardtii***

2.3.2.1 ***Biolistic transformation***

Chloroplast transformation of the *C. reinhardtii* recipient line bst-same1 involved bombardment of a cell lawn with DNA-coated microparticles (henceforth 'Biolistic' transformation (Finer *et al.*, 1999)), and used the Biorad PDS-1000/He Biolistic Particle Delivery System. Cells were grown in 400 ml of TAP medium to mid-log phase (1-2 x 10^6 cells/ml), centrifuged at 4k x *g* for 5 min, and resuspended in fresh HSM to a concentration of 5 x 10^7 cells/ml. Molten HSM agar (0.5 % agar (w/v)) was cooled to 42 °C, and 8 ml added for every 1 ml of cells. The mixture was immediately poured onto HSM amp^{100} agar plates with 4.5 ml per plate. Plates were left in the dark at room temperature for approximately 30 minutes to allow for setting of the soft agar.

The microcarrier particle coating protocol was derived from those previously described (Cullen *et al.*, 2007; Sanford *et al.*, 1993). Tungsten microparticles (0.7 µm) were prepared in advance by washing in 70 % (v/v) ethanol and rinsing three times in sterile ddH_2O before being suspended in 50 % (v/v) glycerol to a concentration of 60 µg tungsten per ml, and stored at 4 °C. Coating of microcarriers was conducted in batches of 50 µg, sufficient for 6 reactions. 50 µl of the above preparation was added to a 1.8 ml eppendorf tube and vortexed

vigorously for 3 minutes in the presence of 1 µg of plasmid DNA, 50 µl 2.5 M $CaCl_2$ and 20 µl 0.1 M spermidine trihydrochloride. The coated microparticles were washed with 70 % ethanol and re-suspended in 48 µl absolute ethanol. This suspension was split and applied to the centre of 6 sterile macro carriers, and were then allowed to air dry in a laminar flow cabinet.

DNA–coated microparticles were fired at previously prepared plates following the instructions for the PDS-1000/He device, using 1100 PSI rupture disks. The plates were then sealed (with two pin pricks per plate to allow gas exchange) and incubated at 25 °C under constant illumination (~50 µmol/m^2/s). Transformant colonies were observed after 2 - 6 weeks and re-streaked to single colonies on HSM agar plates for three rounds prior to genetic and protein analysis.

2.3.2.2 Glass bead-mediated transformation of the C. reinhardtii *chloroplast*

The glass bead method used for chloroplast transformation is adapted from the original protocol developed by Kindle *et al.* (Kindle *et al.*, 1991). The cell wall-deficient recipient line TN72 was grown up in 400 ml of TAP medium to mid-log phase (1-2 x 10^6 cells/ml), centrifuged at 4k x *g* for 5 min and resuspended in fresh HSM to a concentration of 2 x 10^8 cells/ml. 300 µl aliquots of this cell suspension were transferred to sterile 5 ml test tubes containing 0.3 g 425-600 µm diameter glass beads (Sigma). Plasmid DNA (10 µg per plate) was then added to each tube, before vortexing at top speed for 15 s. Molten HSM agar (0.5 % agar (w/v)) was cooled to 42 °C, 3.5 ml added to each tube, and the mixture immediately poured onto HSM amp^{100} agar plates. The plates were left in the dark at room temperature for approximately 30 minutes to allow for setting of the soft agar, after which the plates were treated as for biolistic transformant plates above.

2.4 Protein analysis

2.4.1 Preparation of total protein extracts

During the course of this thesis reference is made to two distinct modes of protein preparation. For separation by SDS-PAGE, preparations of total cellular protein were made by chemical lysis under denaturing conditions. For applications that required preservation of the structure or activity of the proteins, mechanical

breakage was employed to give a soluble protein preparation. This latter method was also used for quantification purposes.

2.4.1.1 ***Protein preparations for SDS-PAGE***

2.4.1.1.1 Preparations from E. coli

E. coli cells were grown overnight to stationary phase as described above. Optical densities were taken and cultures equalised to an OD of 2.5 at 600 nm. 1 ml culture was centrifuged (2 minutes, 21k x *g*) and resuspended in 600 µl of Solution Ab (0.8 M TrisHCl pH 8.3, 0.2 M sorbitol, 1 % (v/v) β-mercaptoethanol, 1 % (w/v) SDS, 0.08 % (w/v) bromophenol blue). Samples were then either used directly for SDS-PAGE, or split into 50 µl aliquots for storage at -80 °C.

2.4.1.1.2 Preparations from C. reinhardtii

C. reinhardtii cultures were grown to mid-late log phase (approximately 3-6 x 10^6 cells per ml) and the optical density at 750 nm measured. 20 ml samples were centrifuged (4k x *g*), and the pellets resuspended in a volume of Solution A (0.8 M TrisHCl pH 8.3, 0.2 M sorbitol, 1 % (v/v) β-mercaptoethanol, 1 % (w/v) SDS) corresponding to the OD_{750} of the sample in ml. For example, a sample with an OD 750 nm of 0.8 would be resuspended in 800 µl Solution A. Samples were then either used directly for SDS-PAGE, or split into 50 µl aliquots for storage at -80 °C.

2.4.1.2 ***Non-denaturing protein preparations***

2.4.1.2.1 Preparations from E. coli

Cells were grown overnight as described above and equalised to an OD_{600} of 12.5 or 25 (representing a concentration of 5 and 10x equalised culture cell concentration, respectively) by centrifugation for 10 minutes at 4k x *g*. The cell pellet was resuspended in sterile Phosphate Buffered Saline (PBS, 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4). Cells were then broken using the OneShot Cell Disrupter (Constant Cell Disruption Systems) at 30k PSI according to the manufacturer's instructions. The cell lysate was split into 1 ml aliquots and centrifuged (30 minutes, 21k x *g*), following which the supernatant was retained and stored at -80 °C.

2.4.1.2.2 Preparations from C. reinhardtii

2.4.1.2.2.1 Cell disruption

Cells were grown as described above. Cultures were measured for optical density at 750 nm and resuspended in a volume corresponding to the culture OD_{750} multiplied by either the (culture volume)/20 (for a ~10x concentration) or (culture volume)/200 (for a ~100x concentration). The resulting cell suspension was loaded into the OneShot Cell Disrupter in batches of 5 ml and disrupted at 10k PSI for cell wall deficient *C. reinhardtii*, or 20k PSI for cell walled strains. In cases of the 100x culture concentration, the resulting cell lysate was diluted to a 10x equivalent concentration with PBS following disruption to avoid precipitation of protein on centrifugation. Samples were incubated on ice for a minimum of 30 minutes (to allow aggregation of presumably membrane fragments), and centrifuged at 7k x *g* for 30 minutes, 21k x *g* for 10 minutes, or 100k x *g* for 1 hour, for purification, activity assays, or quantification assays, respectively.

2.4.1.2.2.2 Lysis of cell-wall-deficient C. reinhardtii *by freeze/thawing*

Lysis of cell wall-deficient strains by freeze/thawing were conducted in two ways: a single round overnight at -20 °C followed by thawing at room temperature, or three rounds of rapid freezing in liquid nitrogen with rapid thawing in a water bath at 30 °C. In either case *C. reinhardtii* cultures were grown and equalised as described above.

2.4.1.2.2.3 Techniques used in cell breakage comparison assays

In all cases the cell wall-deficient strain TN72 was used. Cells were prepared as described above, and each assay conducted on a 1 ml scale followed by centrifugation (30 minutes, 21k x *g*) and preparation for SDS-PAGE loading by addition of 4x protein loading dye as described below. Osmotic shock breakage was by resuspension of cells in ddH_2O in place of PBS followed by incubation at room temperature for 30 minutes. Breakage due to sheer forces involved vortexing at high speed in a 1.8 ml eppendorf tube for 1 minute. Non-denaturing chemical lysis of cells was achieved by addition of the non-ionic detergent Triton X-100 to a final concentration of 1 % (v/v) and incubating at room temperature for 30 minutes.

2.4.2 Analysis by denaturing polyacrylamide gel electrophoresis (SDS-PAGE)

2.4.2.1 ***Sample preparation***

Sample preparation for SDS-PAGE was conducted in two ways. Chemically broken *C. reinhardtii* cells in Solution A, and *E. coli* cells in Solution Ab, were heated to 98 °C for 2 minutes following which they were briefly cooled on ice, centrifuged (2 minutes, 21k x *g*), and the supernatant retained. Mechanically broken cell samples and samples from the purification process were prepared by the addition of 4x protein loading dye (200 mM Tris-HCl pH 6.8, 8 % (w/v) SDS, 40 % (v/v) glycerol, 4 % (v/v) β-mercaptoethanol, 50 mM EDTA, 0.08 % (w/v) bromophenol blue) to a final 1x concentration. These samples were then treated as for the Solution A based methods.

2.4.2.2 ***SDS-polyacrylamide gel preparation***

SDS-PAGE gels were cast using the Bio-Rad mini-PROTEAN Tetra System. Individual gels were made to the following specification. The resolving gel was made to a final acryl/bis concentration of 15 % (w/v) and consisted of 2.25 ml 40 % (w/v) acrylamide:bisacrylamide at 37:1, 750 µl resolving gel buffer (3 M Tris-HCl, pH 8.8), 60 µl 10 % (w/v) SDS, and 2.94 ml ddH_2O. In order to initiate polymerization 250 µl 10 % (w/v) ammonium persulphate and 2.5 µl TEMED were added and the gel poured immediately with 4 ml and 6 ml used per gel for 1.0 mm and 1.5 mm thickness gels, respectively. The acrylamide solution was overlaid with absolute ethanol to prevent oxygen related inhibition of polymerization and to give a level transition for the stacking gel. Polymerization was allowed to continue for a minimum of 30 minutes, after which the ethanol overlay was removed and the top of the gel washed several times with dH_2O to remove any un-polymerized acrylamide.

The stacking gel was made to a final acryl/bis concentration of 3.75 % (w/v) and consisted of 198 µl 40 % (w/v) acrylamide:bisacrylamide at 37:1, 523 µl stacking gel buffer (0.5 M Tris-HCl, pH 6.8), 21 µl 10 % (w/v) SDS and 1.26 ml distilled H_2O. Polymerization was initiated with 175 µl ammonium persulphate and 2.5 µl TEMED, and the gel poured immediately with 1.5 ml and 2 ml used per gel for 1 mm and 1.5 mm thickness gels, respectively. The appropriate comb was inserted

into the stacking gel and polymerization allowed to continue for a minimum of 30 minutes before assembly of the gasket and loading of samples.

2.4.2.3 ***Sample loading and running***

25 µl samples were loaded unless otherwise stated, and gels were run in pre-chilled Tris-glycine electrophoresis buffer (0.25 M Tris, 1.92 M glycine, 1 % (w/v) SDS pH 8.3) at 150 V for approximately 2 hours, or until the dye front ran off the bottom of the gel. Gels were then either electroblotted onto nitrocellulose membranes for immunodetection or directly visualized by staining with Coomassie Brilliant Blue R.

2.4.2.4 ***Coomassie Brilliant Blue R staining***

Gels were soaked in Coomassie Brilliant Blue R solution (3 mM Coomassie brilliant blue R, 50 % (v/v) methanol, 10 % (v/v) glacial acetic acid) for 1 hour at room temperature with gentle agitation (30 rpm, Stuart Mini Orbital Shaker SSM1). Stained gels were then rinsed and soaked in destaining solution (40 % (v/v) methanol, 10 % (v/v) glacial acetic acid) with gentle agitation (30 rpm) until all non protein-bound stain was removed, replacing the destaining solution as necessary. Gels were then either photographed or scanned using the LiCor Odyssey scanner at 800 nm.

2.4.3 **Analysis by western blot analysis**

Proteins separated by SDS-PAGE as described above were transferred to Hybond-ECL nitrocellulose membranes (GE Healthcare) by semi-dry electrophoresis. Before transfer, gels were soaked in Towbin buffer (25 mM Tris, 192 mM glycine and 20 % (v/v) methanol) at room temperature for 30 min. Transfer was conducted using the Bio-Rad Trans-Blot SD semi-dry electrophoretic transfer system according to the manufacturer's instructions. The transfer stack consisted of 6 sheets of appropriately sized 3 MM Whatman paper above and below the nitrocellulose membrane (all soaked in Towbin buffer for 10 min) with the gel being placed above the nitrocellulose membrane and below the upper stack of 3 MM Whatman paper. During assembly of the stack all air bubbles were rolled out and excess buffer removed from the electroblotter. Transfer was carried out at a constant voltage of 25 V (Fisons FEC 570 powerpac) for 1 hour. To ensure proteins

were successfully transferred, gels were subsequently stained with Coomassie Brilliant Blue R as described above.

2.4.3.1 ***Immuno-detection***

Following transfer, the membrane was treated to a simultaneous blocking (5 % (w/v) skimmed milk powder in TBS-T (20 mM Tris base, pH adjusted to 7.4 with 5 M HCl, 137 mM NaCl, 0.1 % (v/v) Tween-20)) and primary antibody treatment (rabbit anti-HA, 1:2000) either overnight at 4 °C or for 1 hour at room temperature, in both cases shaking at 60 rpm. The membrane was then washed in TBS-T wash buffer (two brief rinses followed by three rounds of 5 min washes with shaking at 200 rpm). The membrane was then incubated with the secondary antibody (anti-rabbit IgG: horseradish peroxidase linked whole antibody for ECL detection or IRDye 800 nm fluorophore linked whole antibody for LiCor Odyssey detection, diluted in TBS-T with 2.5 % (w/v) skimmed milk powder to an antibody concentration of 1:10,000 and 1:20,000, respectively) for 1 hour at room temperature with shaking at 60 rpm. The membrane was then washed as for the primary antibody washes. For membranes to be scanned with the LiCor Odyssey system, three additional rinses with TBS were conducted to remove any remaining Tween-20. For detection using ECL (SuperSignal West Pico Chemiluminescent substrate, supplied by Pierce) excess TBS-T was removed from the membrane and an equal volume of detection solutions 1 and 2 were mixed allowing a sufficient total volume to cover the membrane. The membrane was incubated with the reagents for 5 minutes and then excess liquid removed. The blot was sealed in between polythene sheets using a heat sealer and placed in an exposure cassette. Hyperfilm ECL (GE Healthcare) was then exposed to the membrane for 30 seconds and developed using an automatic film processor. More exposures of varied lengths were conducted as necessary to a maximum duration of 20 minutes.

LiCor Odyssey detection was conducted on a Licor Odyssey Infrared Imaging System at an intensity level of 5. Membranes were placed wet on to the scanner bed with all air bubbles being removed with a rubber roller. Image analysis and processing was conducted using the Odyssey Application Software version 3.0.21.

2.4.3.2 ***Quantification recombinant protein***

2.4.3.2.1 Quantification of HA tagged protein

Quantification of HA tagged proteins was conducted by western blot comparison with a commercial HA tagged protein (human CARHSP1, AbCam) of known concentration. *C. reinhardtii* samples to be assayed for HA-protein content were prepared and equalised as described above. Prior to loading, a dilution series was made up with Solution A giving final sample concentrations of 0.1x 0.25x, 0.5x, 0.75x and 1x, where 1x refers to standard equalised *C. reinhardtii*. CarHSP1 samples were made up to 100, 200 and 500 ng per 25 µl aliquot. SDS-PAGE followed by western blot and immunodetection was conducted with visualisation and quantification of bands using the LiCor Odyssey system.

2.4.3.2.2 Quantification of total soluble protein

Quantification of total soluble protein was required in order to calculate protein yield in the form of % Total Soluble Protein (TSP). Samples were prepared to an equalised 10x concentration for mechanical cell breakage as previously described. Cell lysate was then subjected to ultra centrifugation (1 hour, 100k x *g*), and the supernatant analysed for total protein content by the Bradford assay (BioRad) following the manufacturer's instructions. Measurements were taken in triplicate and compared to a calibrated BSA standard curve (Chart 2.1).

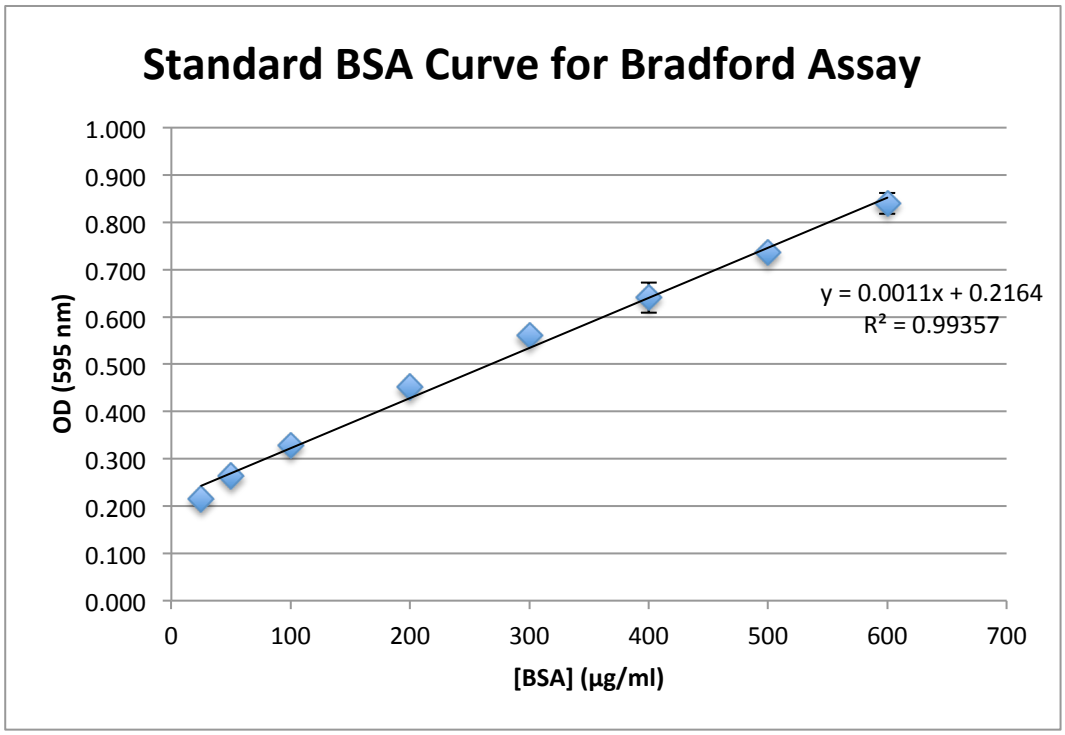


Chart 2.1 – BSA standard curve for Bradford assay

Calibration curve for the Bradford assay created using Bovine Serum Albumin at concentrations ranging from 25-600 µg/ml assayed at 595 nm

2.4.4 Protein stability analysis

C. reinhardtii cells for stability assays were prepared to an equalised 10x concentration and broken by cell disruption as previously described. To represent a truly crude cell extract the resulting cell lysate was not centrifuged however chloramphenicol was added to a final concentration of 25 µg/ml to prevent any continued translation. Samples were incubated in total darkness at 4 °C, 25 °C or 37 °C with 60 µl samples being taken at regular intervals and immediately mixed with 20 µL 4x protein loading dye and frozen at -80 °C. After the final sample had been taken all samples were thawed and analysed by western blot as described above.

2.5 Protein purification

Cpl-1 and the Cpl-1 containing fusion proteins were purified by DEAE ion exchange in a protocol adapted from Loeffler *et al.* (Loeffler *et al.*, 2001) using free choline as a specific elutant. Purification was conducted at 4 °C. Buffers used in the purification protocol are listed in Table 2.4.

The column was packed with DEAE-cellulose fast flow fibres (Sigma), having first been soaked in wash buffer 1 for two hours at room temperature. The column was equilibrated with a further two column volumes of wash buffer 1. *C. reinhardtii* cell extracts were prepared as described earlier (2.4.1.2.2.1) to an equalised concentration of 10x cell culture. Samples were loaded to the column by a peristaltic pump (MasterFlex C/L Compact) at a rate of 2 ml per minute. The column was washed with three column volumes of wash buffer 1, followed by four volumes of wash buffer 2, and two volumes of wash buffer 3. Cpl-1 was eluted in two volumes of elution buffer. During loading and wash stages 1/5 column volume fractions were taken by automatic fraction collection (Pharmacia GradiFrac). For elution fractions, the size was reduced to 1/10 column volume unless otherwise stated. Collected fractions were assayed for total protein content using the Bradford technique in 100 µl reactions, with qualitative results being taken by eye. Specific presence of the target protein was detected by one of three methods. Early purification runs were analysed by anti-HA western blot as previously described. Later runs were analysed by anti-HA dot-blot (2 µl spots directly onto

nitrocellulose, immunodetection as described above); however, this method was only applicable to elution fractions due to the high levels of chlorophyll seen in the flow through and wash fractions.

Chlorophyll binds strongly to nitrocellulose, inhibiting the ECL chemiluminescent reaction and fluorescing in the near infrared region making detection of immuno-tagged protein impossible by either detection method. To counter this a new method was developed where chlorophyll containing fractions were spotted onto filter paper as opposed to directly onto nitrocellulose. The paper was then washed in 100 % acetone to remove chlorophyll and other organic pigments while precipitating the protein onto the filter. Protein spots were then transferred onto nitrocellulose by capillary blotting and immuno-detection conducted as described above.

Table 2.4 – Buffers used in DEAE cellulose ion exchange purification of Cpl-1 and related proteins

Protocol adapted from (Loeffler *et al.*, 2001)

Buffer	Make up
Wash 1	20 mM phosphate buffer, pH 7.4
Wash 2	20 mM phosphate buffer, pH 7.4 1 M NaCl
Wash 3	20 mM phosphate buffer, pH 7.4 0.1 M NaCl
Elution	20 mM phosphate buffer, pH 7.4 0.1 M NaCl 6.5 % (w/v) choline chloride

Once elution fractions were identified, those most rich in the target protein were pooled and concentrated using 30 kDa exclusion spin columns (Vivaspin 6, Sartorius Stedim Biotech) to a final volume of 3 ml. Samples were then dialysed in dialysis cassettes (Slide-A-Lyzer G2, Thermo Scientific) against a 100 fold volume of PBS to remove bound choline thus restoring enzymatic activity.

2.6 Lysin activity analysis

2.6.1 Preparation of lysin extracts

2.6.1.1 *Extracts from* E. coli

E. coli protein extracts were prepared by mechanical cell breakage at an equalised concentration of OD_{600} of 12.5 or 25 (approximately a 5x and 10x concentration respectively, relative to a stationary phase culture), as previously described. Samples were centrifuged (10 min, 21k x *g*) and stored as 0.5 ml aliquots at -80 °C.

2.6.1.2 *Extracts from* C. reinhardtii

C. reinhardtii protein extracts were prepared by mechanical cell breakage at an equalised concentration of OD_{750} of 20 (approximately a 10x concentration relative to a mid-late log culture), as previously described. Samples were centrifuged (10 minutes, 21k x *g*) and stored as 0.5 ml aliquots at -80 °C. Purified extractions were prepared as described above and stored as 0.5 ml aliquots at -80 °C.

2.6.2 Preparation of bacterial suspensions

Bacterial cultures for lysis assays were grown to stationary phase overnight. Cultures were measured for optical density then centrifuged (10 minutes, 4k x *g*) and resuspended in PBS to an OD_{600} of approximately 1. Samples were used immediately.

2.6.3 Reaction conditions

2.6.3.1 *Solid media inhibition assays*

Overnight bacterial cultures were spread onto solid medium and allowed to air dry. Samples to be assayed were then spotted (20 µl) directly onto the plate and

allowed to dry prior to overnight incubation as described above. Alternatively sample and bacterial culture were mixed in a one to one ratio and incubated for 30 minutes at room temperature. 20 µl samples were then spotted onto solid medium and incubated as above.

*2.6.3.2 **Liquid clearance assays***

Initial clearance assays were conducted in 1 ml plastic cuvettes (Fisherbrand) at 37 °C. Bacterial suspensions were pre-warmed for 15 minutes prior to the start of the experiment, and kept in insulated polystyrene cuvette holders to prevent cooling when removed from the incubator for measurements. Cuvettes were sealed with Parafilm to prevent evaporation during the course of the experiment. Subsequent assays were conducted on a 100 µl scale in 96 well plates. These were assayed at 37 °C in an EL_X808 IU microplate reader (Bio-tek Instruments Inc.). Lysin extracts were assayed at 5, 10 and 20 % of total reaction volume.

2.7 Bioinformatics analysis

All bioinformatics investigation was conducted on Apple Macintosh based machines running OSX version 10.8. Codon- and codon pair usage data was compiled using the Codon Usage Optimizer v0.9[6]. Analysis of codon based data was conducted in Microsoft Excel for Mac. Local alignment searches were conducted using either pBlast, pBlastn, or Blastn tools[7] (Altschul *et al.*, 1990) using all standard parameters with an E value cutoff of 1×10^{-5}. Basic molecular biology modeling including plasmid mapping, endonuclease digest predictions, and simple alignments was conducted using MacVector v12. Protein structure modeling was conducted using Jmol[8], with secondary and tertiary structure predictions made using the SCRATCH Protein Predictor web interface[9]. Endogenous sequences, including those of *C. reinhardtii* genes were retrieved from the NCBI database[10]. Sequences of previously expressed recombinant genes from the Purton lab were collected from lab members.

---

[6] www.codonusageoptimizer.org

[7] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[8] http://jmol.sourceforge.net/

[9] http://scratch.proteomics.ics.uci.edu/

[10] http://www.ncbi.nlm.nih.gov/pubmed

Chapter 3

The expression of the *S. pneumoniae* specific endolysin, Cpl-1

3.1 Introduction to Cpl-1

Cpl-1, the endolysin of the *Streptococcus pneumoniae* phage cp-1, is arguably the most comprehensively studied of all the known phage endolysins, and hence is a logical place to begin proof of concept investigations into lysin expression in *C. reinhardtii*. It is also a valid choice in its own right; the target pathogen, *S. pneumoniae*, is estimated by the WHO to cause 1.6 million deaths annually, especially in the very young and very old, and is predicted to be responsible for approximately 11 % of global deaths under the age of 59 months (O'Brien *et al.*, 2009). Several drug resistant variants have appeared in recent years making treatment increasingly difficult, and in the context of the ageing population seen in much of the western world, this is predicted to become an increasingly large burden on healthcare systems (Blasi *et al.*, 2012).

Additionally, the nature of typical *S. pneumoniae* presentation lends itself to treatment with a protein-based therapeutic. Although *S. pneumoniae* is responsible for a wide variety of infections, not least meningitis and septicaemia, the infection of the lungs and respiratory tract remains the most widely reported presentation of the disease. Pneumonial presentation allows for direct treatment of the infection by aerosol administration, sidestepping issues surrounding rapid immune clearance of foreign protein associated with, for example, intravenous treatment. Administration to the lungs via an aerosol also allows for demonstration of efficacy of lysins on mucosal membrane infection, and against non-dividing cells.

3.1.1 History of Cpl-1

The *S. pneumoniae* phage cp-1 was first characterised in 1981 and named for the people of the Alcala de Hanares region of Spain, known as 'Complutense', from whom the phage was first isolated. It was noted for its small size relative to similar phage, unique irregular head morphology, and short tail. The particle was determined to consist of nine polypeptides, one of which, with a molecular weight of 39,000 Da, is now known to be the phage lysin Cpl-1; however, no comment was made as to the nature of the protein at this time (Ronda *et al.*, 1981).

In 1987 Garcia *et al* postulated that the 39 kDa protein was a lysin, although analysis of extracts from lysed cells could not conclusively prove this due to the presence of host cell wall degrading enzymes. To resolve this issue, the putative gene was cloned, purified, and shown to indeed possess cell wall degrading activity. It was characterised as a muramidase, cleaving between the carbohydrate moieties of the peptidoglycan backbone, and found to be active only in the presence of choline in the teichoic acid of the *pneumococcal* cell wall. Although with hindsight the antimicrobial potential of such an enzyme is evident, at the time no mention of such was made (Garcia *et al.*, 1987).

3.1.2 Characterisation as a potential next generation antibiotic

The Cpl-1 lysin was proposed as a possible novel antibiotic in 2003 by the Laboratory of Bacterial Pathology and Immunology at the Rockefeller University, New York, headed by Prof Vincent Fischetti. Purified recombinant Cpl-1 was found to be extremely active against *S. pneumoniae* in mouse models, effectively clearing the nasopharyngeal mucosa via a topical application, and dramatically reducing bacterial titre (≥99 %) in the bloodstream. Investigations also showed the half-life of Cpl-1 in the mouse bloodstream to be 20.5 minutes, and produced encouraging results in immunogenic studies and general long-term stability of the protein. Crucially, the presence of the Cpl-1 lysin showed no toxic effect on the mice used in this trial (Loeffler *et al.*, 2003). Further reports from this lab have shown synergistic activity when Cpl-1 is combined with penicillin or gentamicin (Djurkovic *et al.* 2005); effective treatment of endocarditis and pneumococcal meningitis in rats (Entenza *et al.*, 2005; Grandgirard *et al.*, 2008), and the use of Cpl-1 to rescue mice with fatal pneumococcal pneumonia (Witzenrath *et al.*, 2009). In each case the use of Cpl-1 was shown to be highly effective in the treatment of the bacterial infections.

3.1.3 Molecular characterisation of Cpl-1

Over the last decade several structural studies have been conducted on Cpl-1, and the mechanism of action derived. The protein has the bi-modular domain structure common to many of the lysins (as described, 1.1.4.4), and also shows homology with the *S. pneumoniae* autolysin, LytC. Interestingly, the cell wall binding domain

of the native host protein shows greatly reduced stability compared to its phage counterpart, which is thought to be important in prevention of autolysis of the host cell (Monterroso *et al.*, 2008).

Cpl-1 remains one of the few complete lysins with a known crystal structure, shown in Figure 3.1. The catalytic N-terminal domain of Cpl-1 (shown in green) displays an irregular β-barrel motif comprised of five β-sheet/ α-helix hairpins and three stand alone β-strands (known as a $(\beta/\alpha)_5\beta_3$ configuration). The active site forms a long groove at the C-terminal end of the catalytic barrel and conforms to previously described peptidoglycan binding sites. Catalysis is thought to take place via a general acid/base reaction involving Asp10 as the primary nucleophile, and Gln94 as the proton donor (Hermoso *et al.*, 2003).

The cell wall binding domain (CBD) is formed of two distinct structural regions, CI (blue), and CII (magenta), containing six choline binding repeats. Of the four putative choline-binding sites formed, only two are thought to be active. These are both located on CI and can be seen with choline bound in Figure 3.1. The second structural region, CII follows a β-sheet structure and is absent in LytC. It is thought to be involved in interactions between the two domains increasing stability of the CBD (Hermoso *et al.*, 2003). Upon the binding of choline, CBD mediated dimerisation is thought to occur, lending further stability to the complex. This has been shown to be important for the positioning of the catalytic domain relative to the substrate, as demonstrated by removal of the CBD resulting in a drop of catalytic activity by five orders of magnitude (Monterroso *et al.*, 2008).

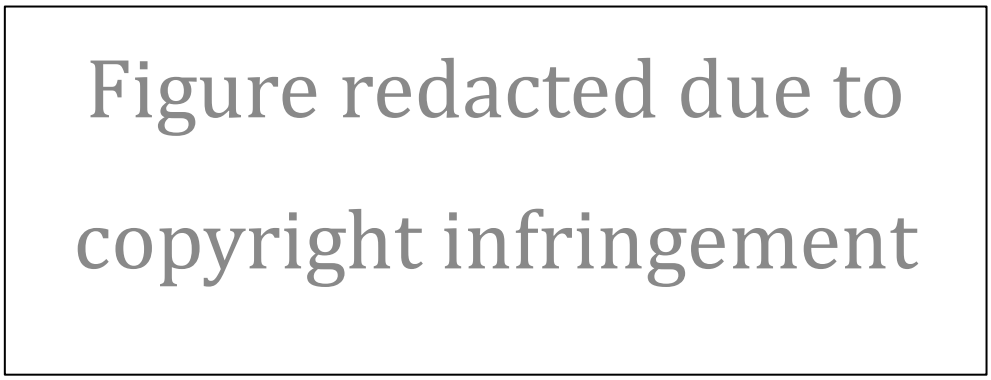


Figure 3.1 – Ribbon representation of the crystal structure of Cpl-1

The crystal structure of Cpl-1 shows an N-terminal catalytic (green) and C-terminal cell wall binding domains (blue and magenta) connected by a flexible acidic linker. Choline molecules are shown as ball and stick representations bound to the CBD. Reproduced from (Hermoso *et al.*, 2003).

3.1.4 Suitability of *C. reinhardtii* for *cpl-1* expression

Considerable work has already been conducted on Cpl-1, and it has been shown to hold great potential as a novel protein-based antibacterial. As with all such therapeutics, cost of production is an issue, with the bulk of the expense involved in production of a usable product being the high levels of purification required to remove the highly immunogenic endotoxins found in *E. coli*. In order for any novel therapeutic to reach the market minimisation of costs is important, but with novel anti-bacterials this can be seen as even more crucial given the perceived risks to profits from acquired resistance, short treatment courses, and humanitarian pressures to supply treatment to the developing world. As discussed in Chapter one, the expression of such a product in the *C. reinhardtii* chloroplast could help to reduce costs of protein based therapeutics, not least due to low production costs and by allowing for less rigorous purification on account of the organism's GRAS status. Clearly such work is still at a very early stage and getting Cpl-1 into the clinic can be seen as a distant goal. This project is proposed as a proof of concept, both for the expression of lysins and indeed therapeutic proteins in general in the *C. reinhardtii* chloroplast.

3.1.5 Aims and objectives

1. To create transgenic lines of *C. reinhardtii* in which the *cpl-1* gene is introduced into a specified locus of the chloroplast genome.

2. To show expression of *cpl-1* in the *C. reinhardtii* chloroplast.

3. To confirm correct folding and absence of deleterious post-translational modification by demonstration of activity against *S. pneumoniae in vitro*.

4. To develop a viable purification strategy for the production of an enriched Cpl-1 preparation.

3.2 Results

3.2.1 The generation of transformant lines containing the *cpl-1* gene

In total three *cpl-1* expressing lines of *C. reinhardtii* were created. Each contained the *cpl-1* gene in its synthetic form, but differed in cellular background and expression cassette construction. An illustration of cell line naming protocols can be seen in Figure 3.2.

3.2.1.1 *Design of a synthetic codon optimised version of* cpl-1

The first *cpl-1* expressing line of *C. reinhardtii* was produced using the expression plasmid pASap1 (Figure 3.3) and the recipient cell line bst-same1. The *cpl-1* gene sequence was retrieved from Genbank (accession number: NP_044837, from the *S. pneumoniae* phage CP-1, NC_001825) and synthesised by GeneArt. The gene was codon optimised to a Codon Adaptation Index (CAI) of 0.8 using the Kazusa CAI table; in total 219 bases were amended giving the optimised and native DNA sequences 79 % identity. In addition to optimisation, a 5' *Sap*I and 3' *Sph*I site was added, as was a downstream Haemagglutinin (HA) epitope tag preceding a double TAA stop codon. With all modifications the expected molecular weight of *cpl-1* gene product was increased to 40 kDa.

The *cpl-1* gene was excised from the GeneArt stock pMK plasmid and inserted into pASap1 following the cloning strategy shown in Figure 3.4 to create the transformation vector pASap1.cpl-1. Confirmation of insertion was by PCR (Appendix a) and DNA sequencing.

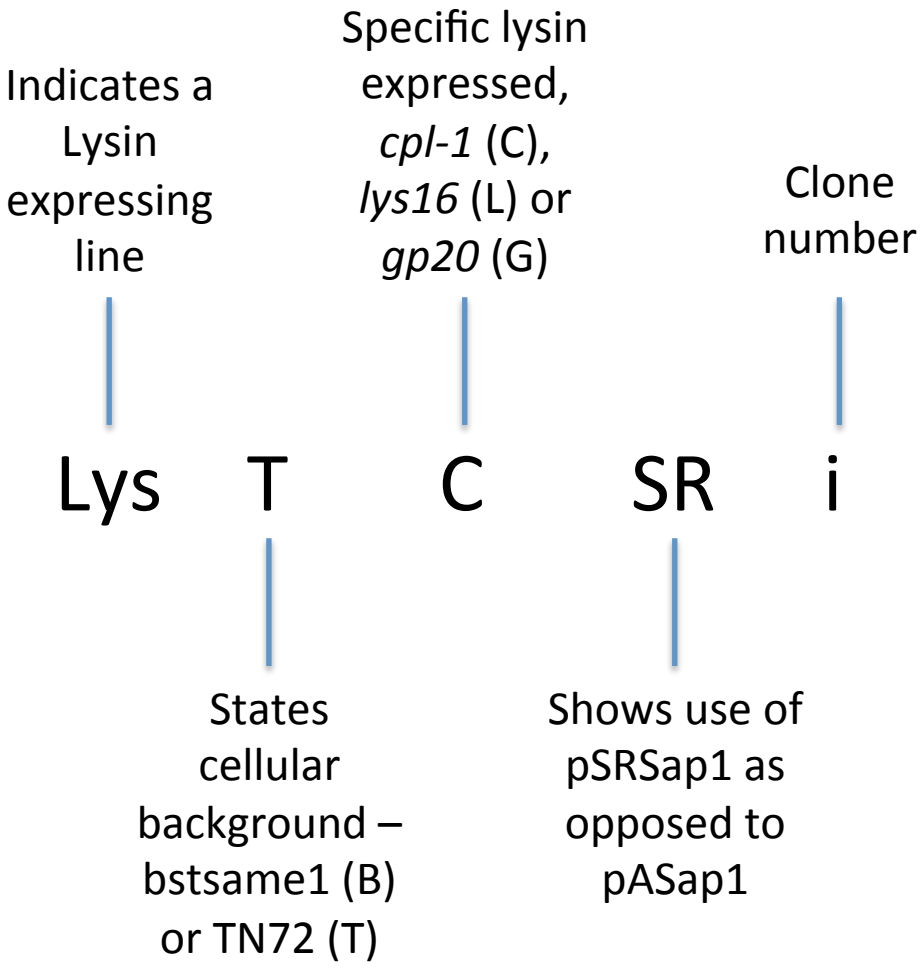


Figure 3.2 – Naming protocols for lysin expressing cell lines

In an effort to create a logical naming system for the various lysin-containing strains, the above general formula was derived. As an extension to this practice, fusion constructs are denoted with the use of a colon, e.g. LysTC:P for a *cpl-1:pal* fusion in TN72.

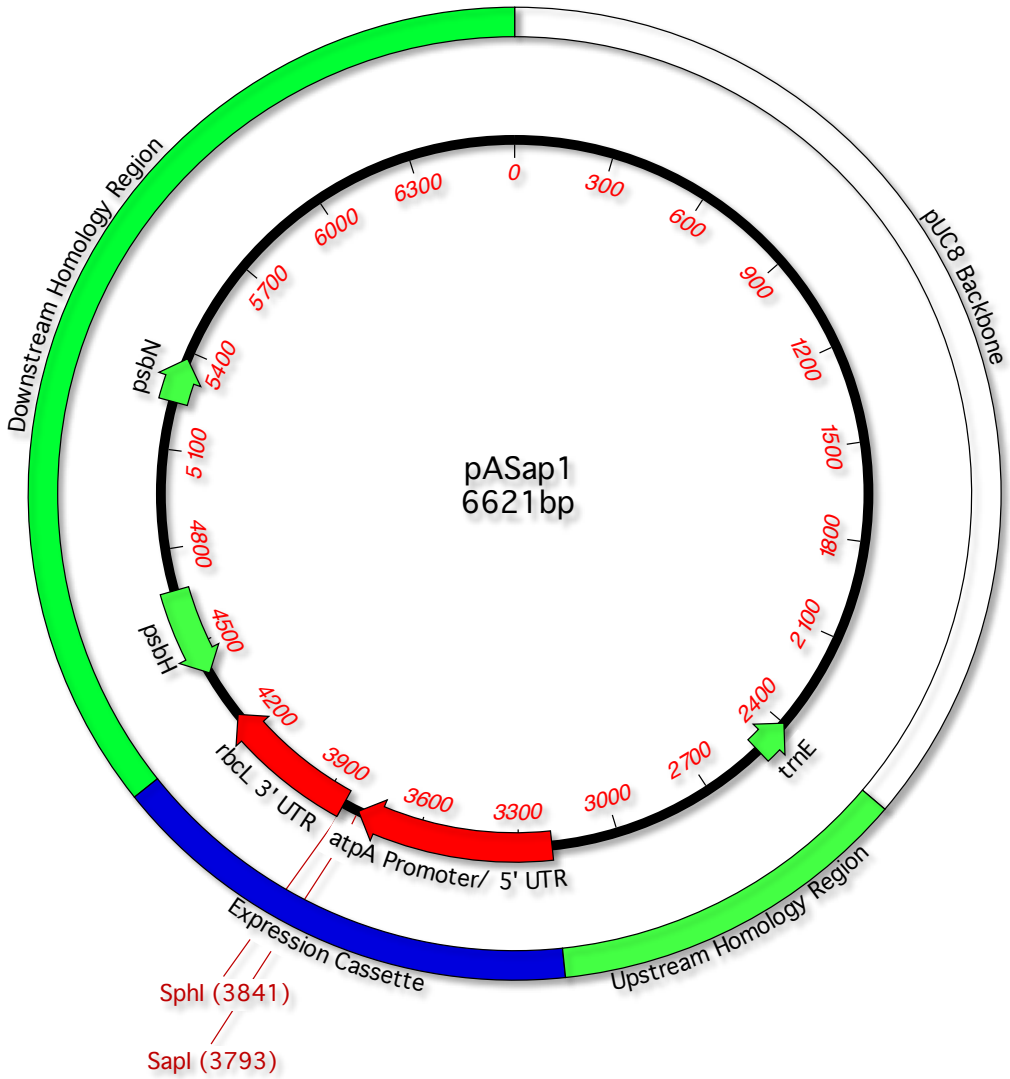


Figure 3.3 - The pASap1 *C. reinhardtii* expression vector

The pASap1 vector is built on a pUC8 backbone and incorporates an expression cassette comprising of the *C. reinhardtii atpA* promoter and 5' UTR, and the *rbcL* terminator and 3' UTR. Flanking the expression cassette are regions homologous to the *C. reinhardtii* chloroplast genome incorporating a functional *psbH* gene for the photosynthetic recovery of Δ*psbH* recipient lines. The expression cassette is targeted to the intergenic region between *psbH* and *trnE*, a neutral locus. The use of *Sap*I as a 5' insertion site allows for a native transition from the 5' UTR into the gene of interest without leaving an endonuclease "scar" as the *Sap*I enzyme cuts outside of its recognition site.

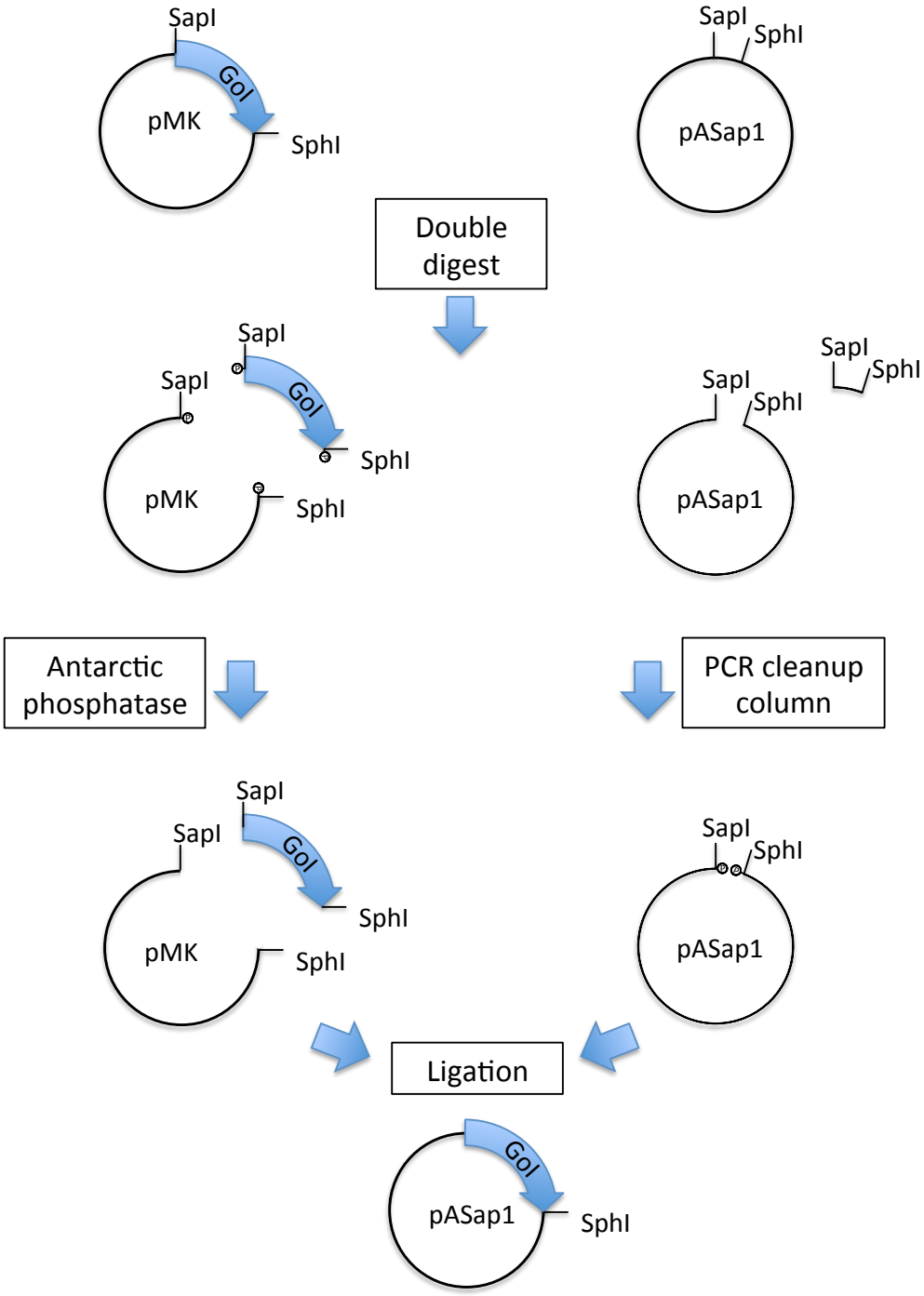


Figure 3.4 - Cloning strategy used to build *C. reinhardtii* transformation plasmids

This general strategy was employed for the construction of all lysin-containing plasmids. The Gene of Interest (GoI) was excised from the stock plasmid by *Sph*I/*Sap*I double digest. The GoI was then dephosphorylated and the excised minor backbone pASap1 fragment removed using a PCR isolation column prior to ligation and transformation into the *E. coli* strain DH5α.

3.2.1.2 Transformation of the C. reinhardtii *strain bst-same1 using pASap1.cpl-1*

The plasmid pASap1.cpl-1 was used to transform bst-same1 via the biolistic method as described (2.3.2.1). Transformation yielded 10-20 colonies per plate, of which 12 were selected for PCR screening (Appendix b). Of the six transformants showing correct insertion of the *cpl-1* expression cassette, two were selected for continued investigation. Correct insertion was confirmed by DNA sequencing and the lines named LysBCi and LysBCii in accordance with the naming protocol shown in Figure 3.2.

3.2.1.3 Transformation of the cell-wall deficient strain TN72 with pASap1.cpl-1

The recipient strain TN72 was designed and created by Dr Thanyanun Ninlayarn, primarily in order to address two issues with the previous recipient line, bst-same1:

- **Disruption of *psbH* gene by point insertion**. Though bst-same1 is suitable as a non-photosynthetic recipient strain, the insertional disruption of *psbH* with the *aadA* cassette does occasionally allow for restoration-only recombination as shown in Figure 3.5. In contrast, in the new strain TN72 the region extending from the middle of *psbH* to the site of insertion of the gene of interest is replaced with the aadA cassette. This arrangement prevents inappropriate recombination and thus averts restoration of the wild type *psbH* locus without the insertion of the GoI.

- **Presence of a cell wall.** Though not a problem for biolistic transformation, the presence of a functional cell wall in bst-same1 prevents transformation via the glass bead method without first removing the cell wall. Although this can be accomplished with the *C. reinhardtii* cell wall degrading enzyme autolysin, the production of natural autolysin from suspensions of mating cells can be unreliable and no commercially available alternative currently exists. TN72 was generated using a cell wall deficient mutant (carrying the *cw15* nuclear mutation) allowing for naïve transformation by glass beads without prior autolysin treatment. The absence of a cell wall in TN72 also allows for significantly easier mechanical cell breakage for non-denaturing harvest of protein.

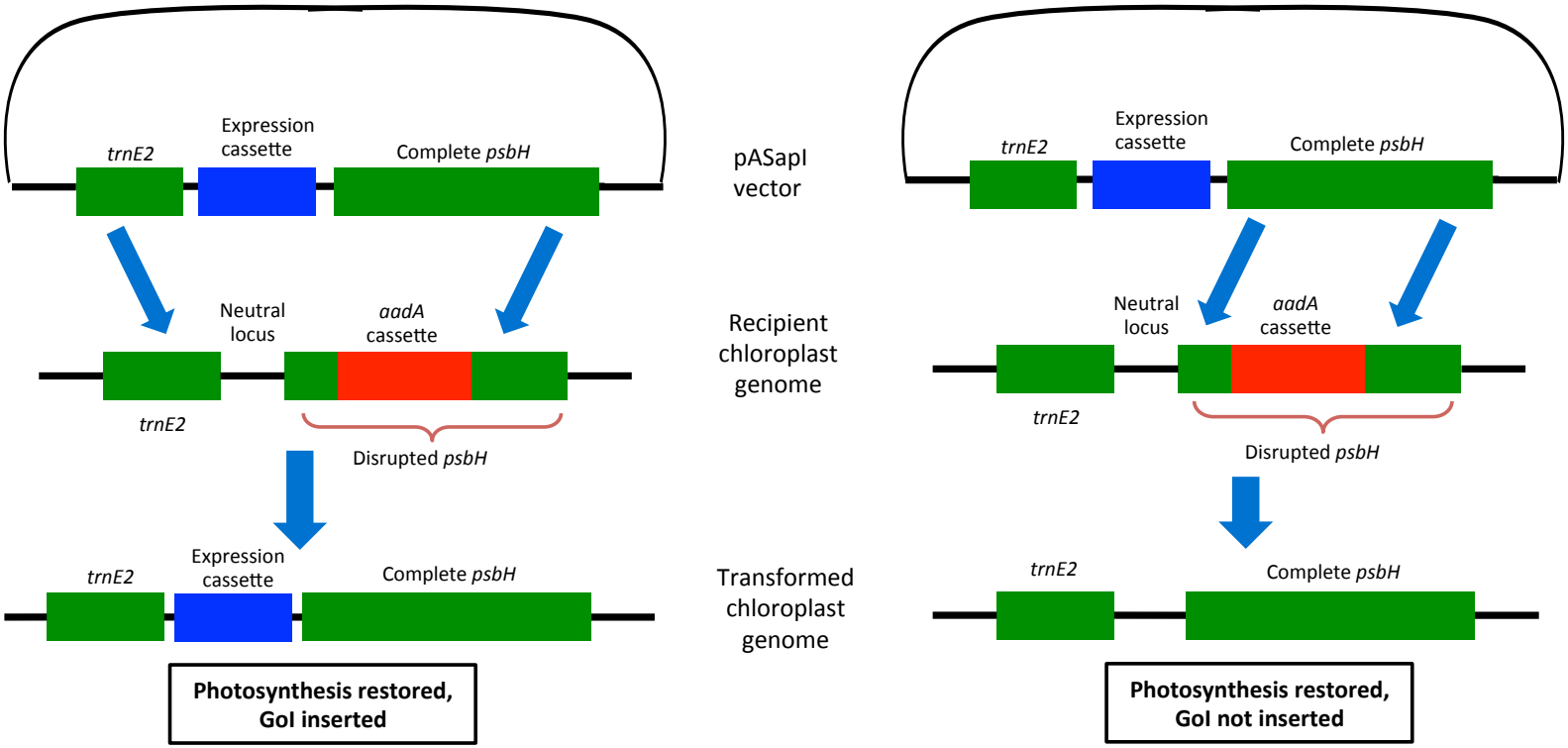


Figure 3.5 – Restoration-only transformation of bst-same1

Left panel: In the desired recombination event, double homologous recombination occurs between the *trnE2* and *psbH* flanking regions resulting in restoration of *psbH* and the insertion of the Gene of Interest. Alternatively the **right panel** shows a double recombination event purely within the *psbH* gene resulting in restoration of *psbH,* but no insertion of the target gene. The recipient strain TN72 avoids this problem via the replacement with the *aadA* cassette of both the *psbH* sequence and upstream genomic region as far as the neutral locus.

Transformation of TN72 was conducted by the glass bead technique using pASap1.cpl-1 as in 3.2.1.1. Transformation yielded low numbers of transformants, with a total of 22 colonies from 12 plates. Two colonies were screened by PCR (Appendix c) and DNA sequencing. On the confirmation of correct insertion of *cpl-1* the resulting strains were named LysTCi and LysTCii.

3.2.1.4 ***Generation of new TN72 transformants using a superior expression plasmid - pSRSap1.cpl-1***

Recent work conducted Dr Thanyanun Ninlayarn and Dr Joanna Szaub in the Purton lab has shown that the expression of transgenes using the *C. reinhardtii psaA* promoter/ 5' UTR gives an approximately two-fold increase in recombinant protein accumulation relative to those driven off of the *atpA* promoter/ 5' UTR elements. Building on this research, a new expression vector designated pSRSap1 was built by Dr Rosie Young by replacement of the *atpA* promoter and 5' UTR of pASap1 with those of *psaA*.

The *cpl-1* gene was cloned into this new vector as for pASap1 (3.2.1.2), and shown to be correct by restriction endonuclease digestion (Appendix d), and confirmed by DNA sequencing. The resulting pSRSap1.cpl-1 plasmid was used to transform TN72 as described above, yielding 15 transformant colonies from 10 plates. Again screening was by PCR (Appendix e) with correct insertion confirmed by DNA sequencing. The resulting strains were named LysTC-SRi and LysTC-SRii.

3.2.2 **Confirmation of expression of *cpl-1* in *E. coli* and *C. reinhardtii***

Due to the shared ancestry of the *C. reinhardtii* chloroplast and the *E. coli* cell, many chloroplast promoters show activity in *E. coli* (Bateman and Purton, 2000; Goldschmidt-Clermont, 1991). Such pre- *C. reinhardtii* expression is used throughout this thesis as a final confirmation of correct plasmid construction, an independent analysis of enzymatic activity, and a means of investigating lysins when successful expression was not achieved in *C. reinhardtii* (as discussed in Chapter four).

Expression of *cpl-1* was confirmed in both expression platforms by western blot analysis with anti-HA antibodies, and visualized by infrared fluorescence with an appropriately labelled secondary antibody (unless otherwise stated). In all cases Cpl-1 ran on gels at an apparent mass of 38-39 kDa, slightly under the predicted 40 kDa. However, it is assumed to be full length because of the self-evident presence of the C-terminal HA tag, and the demonstration of enzymatic activity, as discussed in section 3.2.6.

3.2.2.1 Western blot analysis demonstrates expression of cpl-1 *in* E. coli *under the* atpA *promoter/ 5' UTR*

The pASap1.cpl-1 construct in the *E. coli* strain DH5α gave clear expression of *cpl-1*, as shown in Figure 3.6. The two smaller bands present could be due to either N-terminal degradation products or potentially the result of translation at internal initiation regions. Figure 3.7 shows the predicted regions of truncation in relation to the Cpl-1 protein sequence, highlighting local methionine residues. Although the larger truncated band fragment shows close proximity to a methionine, the smaller fragment does not. Analysis of the nucleotide sequence immediately upstream of the methionine codon shows very little similarity to the *E. coli* consensus Shine Dalgarno sequence, but neither does the *atpA* 5' UTR immediately upstream of the translational start (Figure 3.8). Furthermore, when tracked onto the Cpl-1 crystal structure, both potential cleavage regions are in exposed unstructured surface loops, suggesting easy accessibility to proteases (Figure 3.9). The discrete nature of the degradation products implies endo- as opposed to exo- protease activity. Structure-related cleavage is encouraging as it implies correct folding of Cpl-1 in *E. coli*. This is in agreement with both the prokaryotic origins of the gene product, and reports from other groups. It should be noted that other degradation products might be present; however, where the HA tag has been removed such products would not be detectable by western blot analysis with the currently available antibodies.

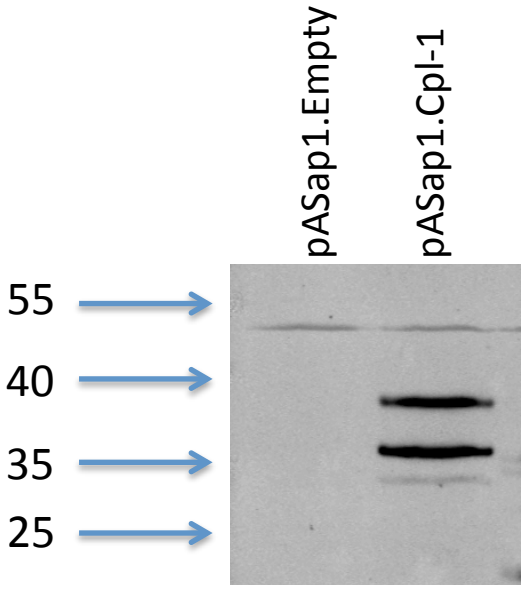


Figure 3.6 – Expression of *cpl-1* in the *E. coli* strain DH5α is confirmed by western blot analysis with anti-HA antibodies

Expression of *cpl-1* is clearly illustrated by a strong band running at an apparent mass of ~39 kDa. The nature of the two smaller bands seen in the pASap1.Cpl-1 line is still in question. The ~52 kDa band seen in both lanes is a non-specific band observed in all *E. coli* western blot analysis with anti-HA antibodies.

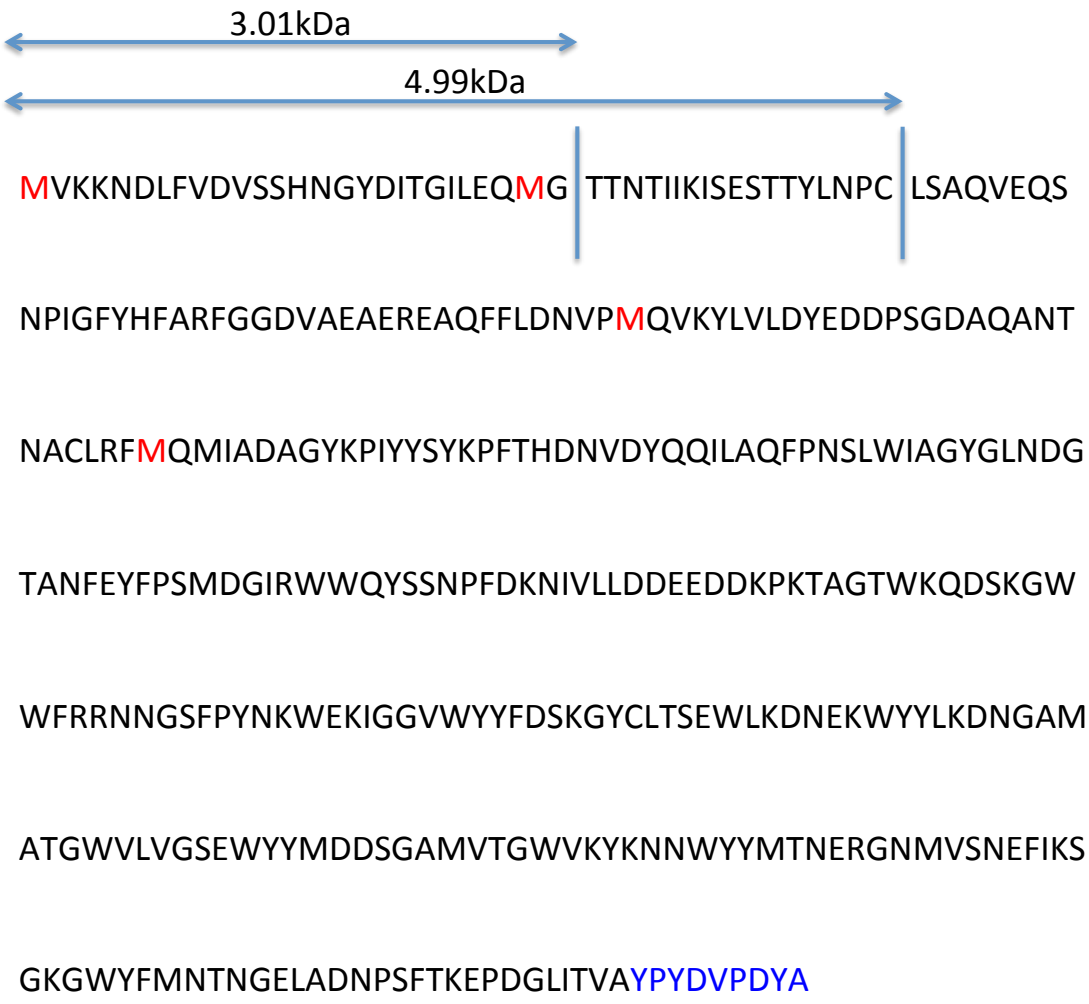


Figure 3.7 – Predicted regions of truncation in relation to the Cpl-1 primary sequence, and relevant methionine residues

Predicted truncation sizes were deduced from Figure 3.6 and mapped onto the Cpl-1 primary sequence. The larger of the two truncation products (~3 kDa deletion) has a methionine in close proximity to its predicted N-terminus suggesting a possible cryptic internal translation initiation region. The smaller of the products has no such methionine. Translation in *E. coli* can also initiate at GUG sites; however, none are seen in frame with the HA tag (shown in blue).

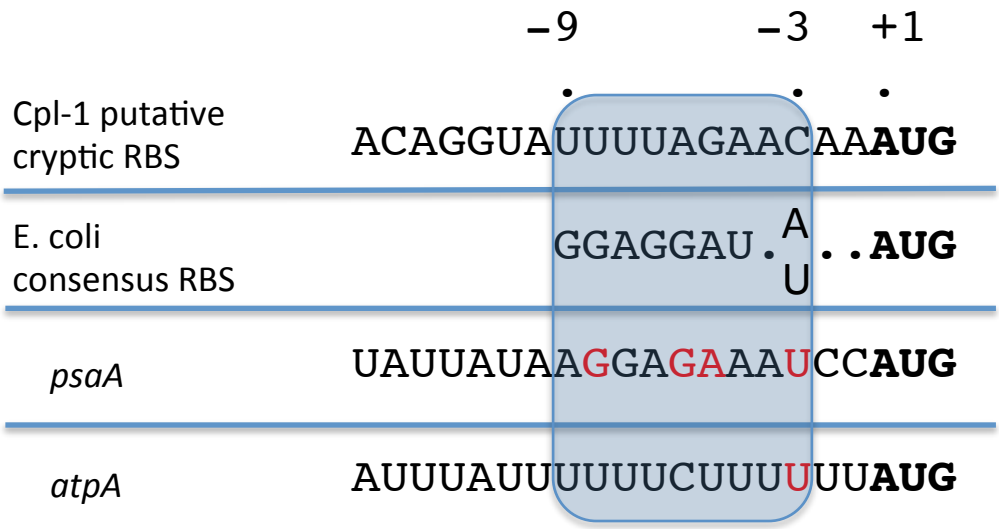


Figure 3.8 – Comparison of a possible cryptic Shine Dalgarno sequence in the *cpl-1* gene with the ideal *E. coli* consensus, and the *atpA* and *psaA* 5' UTRs

The putative cryptic Shine Dalgarno sequence upstream of the large truncation product does not show any homology with the *E. coli* consensus sequence. Neither, however, does the *atpA* 5' UTR, despite being active in this organism.

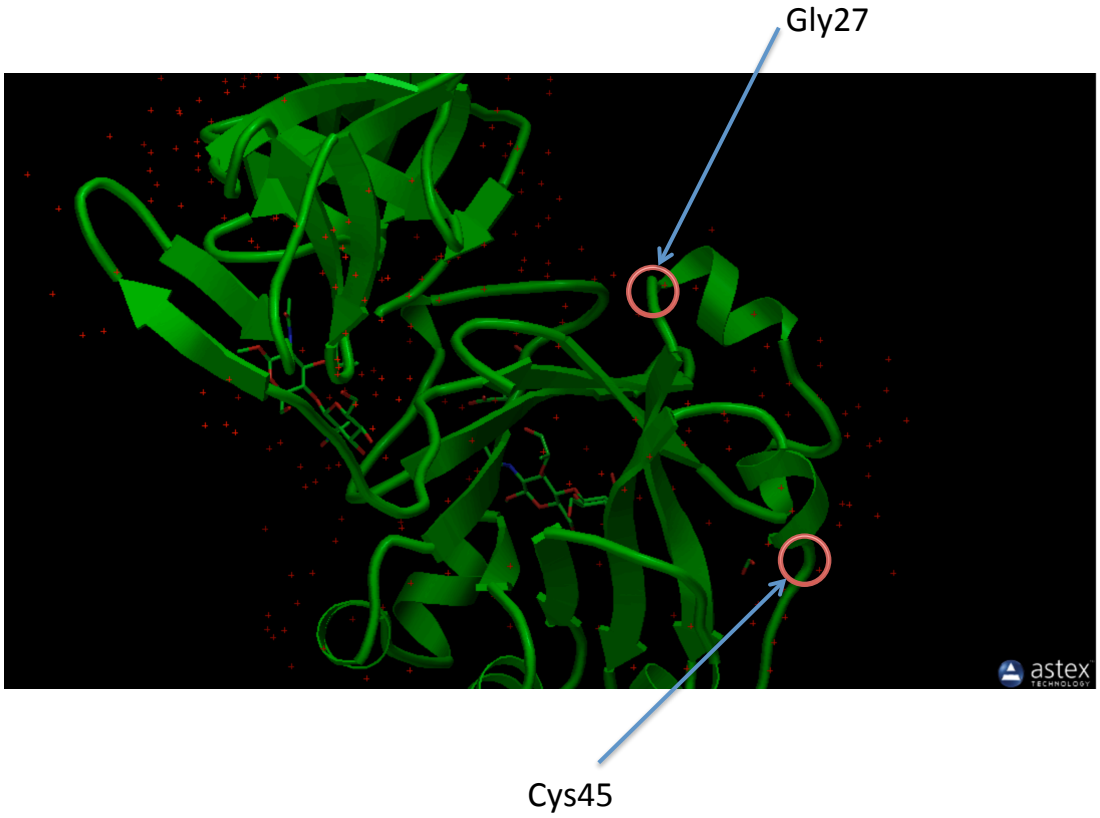


Figure 3.9 – Crystal structure of Cpl-1 with predicted protease cleavage sites

Predicted cleavage sites were mapped onto the crystal structure for Cpl-1[11]. Both sites occur on exposed surface loops supporting the hypothesis that the observed Cpl-1 truncations are a product of site-specific proteolytic degradation.

[11] http://www.rcsb.org/pdb/explore/explore.do?structureId=2J8F

3.2.2.2 *Successful expression of* cpl-1 *in* C. reinhardtii *bst-same1 under the* atpA *promoter/ 5' UTR*

For *C. reinhardtii* to be considered as an expression platform for Cpl-1 it was essential that recombinant protein accumulation could be demonstrated. This was achieved by western blot analysis. The transgenic lines LysBCi and LysBCii were grown to late log phase, harvested, equalised and prepared for SDS-PAGE using the Solution A method as described (2.4.1.1). ECL western blot shows readily detectable levels of full-length Cpl-1 in both lines as shown in Figure 3.10.

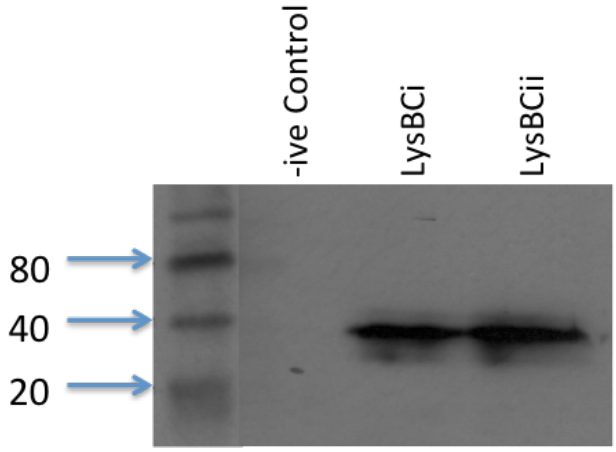


Figure 3.10 – Expression of *cpl-1* in the *C. reinhardtii* bst-same1 is demonstrated by western blot analysis with anti-HA antibodies

Both lines of LysBC (pASap1.cpl-1, bst-same1 background) show strong expression of an HA tagged protein of the correct size which is assumed to be Cpl-1. Wild type *C. reinhardtii* is used as a negative control. This immunoblot was visualised by ECL with a 5-second exposure.

3.2.2.3 **cpl-1** ***is also expressed in the cell wall deficient strain TN72***

Despite the closely related nature of the recipient lines bst-same1 and TN72, and the identical expression cassettes used in each case, TN72 represented a novel cellular background, and thus it was necessary to again ensure Cpl-1 accumulation was occurring. The transgenic lines LysTCi and LysTCii were prepared as described in 3.2.2.2 and expression investigated by western blot analysis. Also included is C6, an example of an aberrant transformant line in which a rare secondary recombination event had occurred between the *atpA* element of the inserted expression cassette and that of the excised *aadA* cassette resulting in the loss of the *cpl-1* gene.

As expected LysTCi and LysTCii also show correct expression of Cpl-1 (Figure 3.11), although it appears to be at a marginally lower level to that seen in the positive control (bst-same1 transformant LysBCii). This has been attributed to generally lower fitness of the cell wall deficient cell lines; however, it could also be an artefact in relation to equalisation of the two different cell lines.

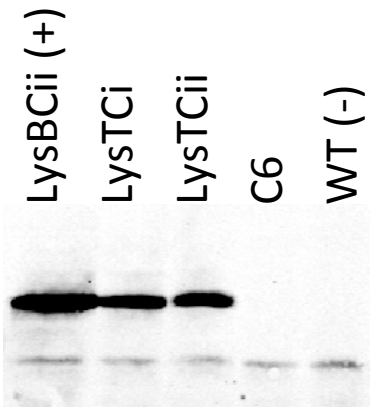


Figure 3.11 – Expression of *cpl-1* in the *C. reinhardtii* LysTC lines is confirmed by western blot analysis with anti-HA antibodies

Expression of *cpl-1* in LysTC is confirmed in both cases, albeit at slightly lower levels than seen in LysBC. Also included is the double recombination line C6, which as predicted does not show expression.

3.2.2.4 The use of the psaA *promoter/ 5' UTR improves* cpl-1 *expression*

The use of the *psaA* promoter/ 5' UTR combination for transgene expression had previously been demonstrated in the *C. reinhardtii* chloroplast (R. Young, unpublished work), but it was necessary to confirm this for the *cpl-1* gene. Expression of *cpl-1* under the *psaA* promoter/ 5' UTR of pSRSap1 was confirmed by western blot analysis (Figure 3.12). In order to directly compare *cpl-1* expression from the *atpA* and *psaA* constructs, dilutions were prepared for LysTCi and LysTC-SRi protein extracts (Figure 3.13,). Despite mis-running of the highest concentration of the LysTC-SRi preparation, quantification of the bands shows an approximate 68 % increase in protein yield when using the *psaA* promoter/5' UTR over that of *atpA*.

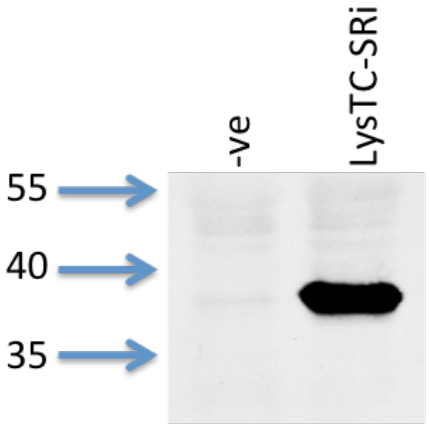


Figure 3.12 – Expression of *cpl-1* in the *C. reinhardtii* lines LysTC-SR is demonstrated by western blot with anti-HA antibodies

A high level of *cpl-1* expression is seen when fused to the *psaA* promoter/5' UTR (although the absence of an *atpA*-driven control prohibits comment on relative accumulation in this case).

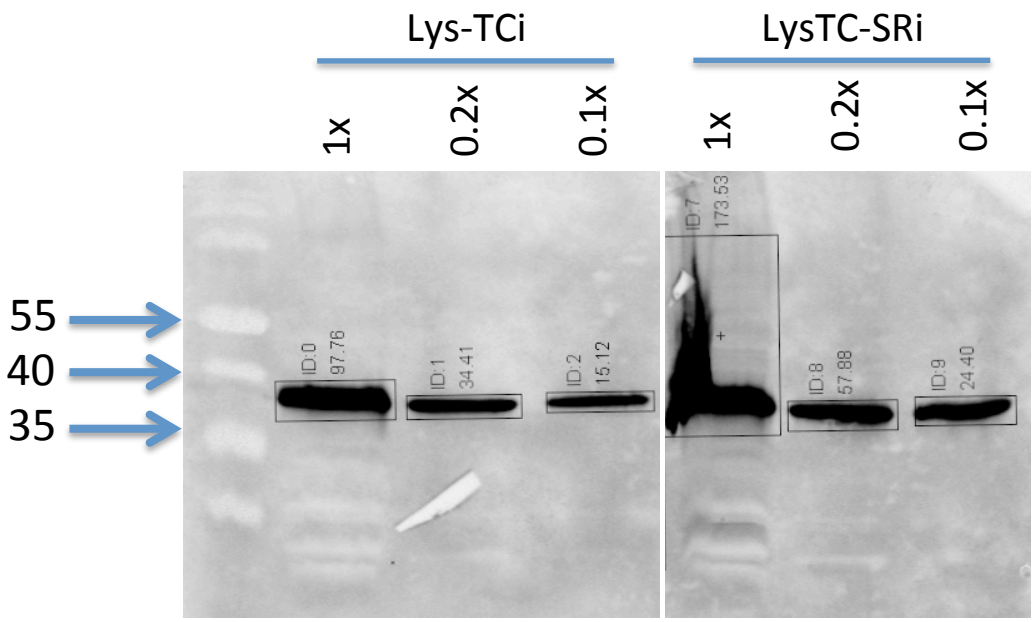


Figure 3.13 – Expression of *cpl-1* in *C. reinhardtii* under the *psaA* promoter/ 5' UTR (LysTC-SRi) is shown to be approximately 1.7-fold that of *atpA* (LysTCi) by western blot analysis with anti-HA antibodies

Dilutions of both lines were blotted onto the same membrane to exclude membrane bias. Despite the poor running of the undiluted LysTC-SRi lane, it is seen that the *psaA* promoter/ 5' UTR combination is significantly better than that of *atpA* (Table 3.1).

Table 3.1 – Quantification based comparison of Cpl-1 accumulation under the *atpA* (LysTCi) and *psaA* (LysTC-SRi) promoters/ 5' UTRs

By producing weighted averages from the three dilutions of each line a rough ratio of 1:1.68 can be calculated, or an increase of 68 % total recombinant protein when the *psaA* element is used relative to *atpA*.

Dilution	Intensity	
	atpA	*psaA*
1.00	97.76	173.53
0.20	34.41	57.88
0.10	15.12	24.4
Weighted Mean	140.3	235.6
Ratio	1	1.68

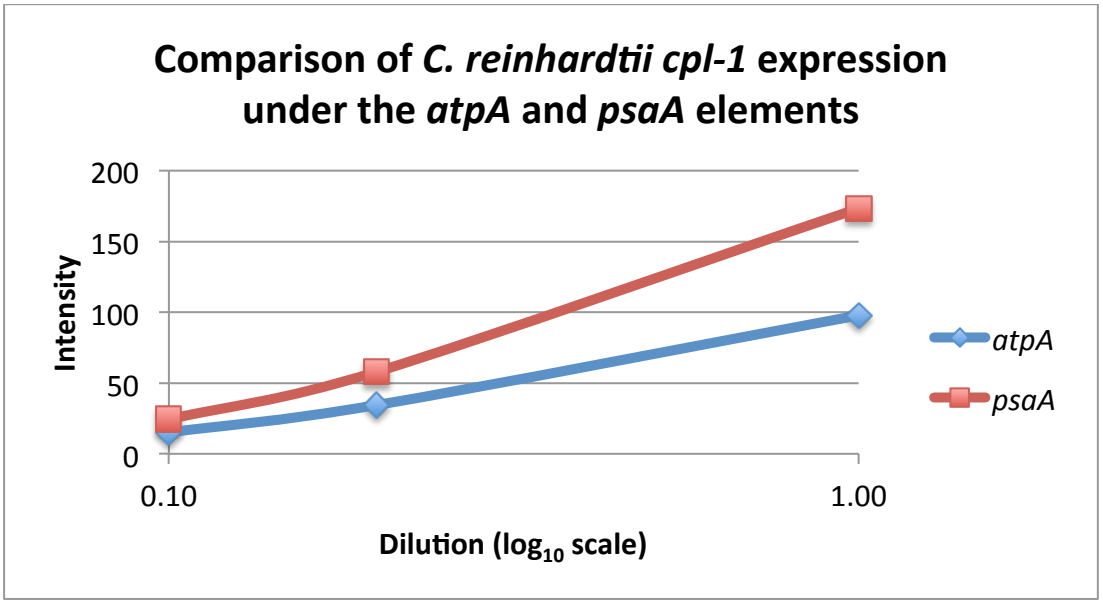


Chart 3.1 – A graphical comparison of *C. reinhardtii* derived Cpl-1 accumulation under the *atpA* and *psaA* promoters/ 5' UTRs

This graphical representation of the data seen in Figure 3.13 shows Cpl-1 yield to be consistently higher when under the *psaA* element relative to that of *atpA.* Expression is measured as a function of secondary antibody fluorophore fluorescence intensity (arbitrary units).

3.2.3 Attempts to quantify Cpl-1 yield and productivity in *C. reinhardtii*

In the literature on recombinant protein production in *C. reinhardtii*, protein yield is generally expressed at a percentage of total soluble protein (%TSP). This unit of measure sidesteps issues related to growth phase and cell density, but is not practical as an illustration of actual productivity as preferred by industry. In order to present data suitable for both academic and industrial audiences yield of Cpl-1 in *C. reinhardtii* has been expressed as %TSP, and also in mg/L for a mid-late log culture (equivalent to an OD of 2 at 750 nm). From the latter measure estimates of mg product/ g dry weight and mg product/ L/ day can be estimated for ideal conditions. In order to present the most highly expressing data only LysTC-SR using the *psaA* promoter/ 5' UTR was considered.

3.2.3.1 *Calculation of Total Soluble Protein by the Bradford assay*

In order to calculate recombinant protein production as a function of %TSP, total soluble protein from equalised mechanically broken samples was measured using the Bradford assay (Bradford, 1976). Soluble protein samples were prepared for both LysTC-SRi and a negative control, BlankT-SR (TN72 transformed with the empty pSRSap1 vector), by mechanical cell breakage followed by ultra centrifugation as described (2.4.1.2.2). Samples were analysed for total soluble protein in triplicate by the Bradford assay calibrated against a bovine serum albumin (BSA) standard curve ranging from 50 – 600 μg/ml (2.4.3.2.2). Total soluble protein in mg/L for equalised 10x cell culture preparations of LysTC-SRi and BlankT-SR is shown in Chart 3.2.

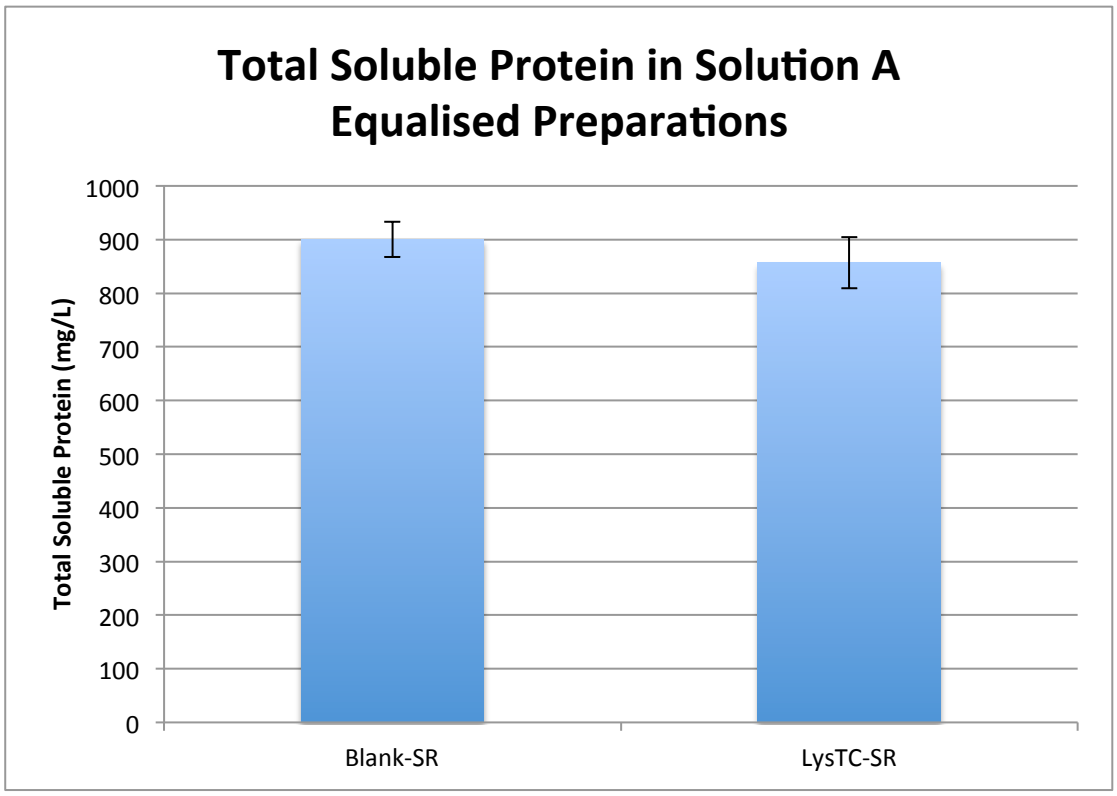


Chart 3.2 - Total soluble protein for *cpl-1* expressing LysTC-SRi and negative control BlankT-SR

Total Soluble Protein (TSP) is displayed for equalised cultures, at 10x cell culture concentration. When scaled to a 1x cell culture concentration these data give 90 +/- 3.3 mg/L and 85.7 +/- 4.7 mg/L for Blank-SR and LysTC-SRi respectively. The slight difference in TSP between the Blank-SR and LysTC-SR is not statistically significant.

3.2.3.2 ***Cpl-1 yield and productivity is calculated by western blot analysis***

Once a value for total soluble protein was acquired, quantification of % TSP could be calculated by measuring the accumulation of Cpl-1 relative to a standard of known concentration. LysTC-SRi samples were prepared for western blot analysis as described (2.4.3) and analysed as a five point dilution series alongside three known concentrations of a commercial HA labelled standard protein, CARHSP. The resulting western blot (Figure 3.14) was quantitatively analysed using the LiCor Odyssey system as previously described and the results are presented in Table 3.2 and Chart 3.3. The data presented gives a %TSP of ~9 %, or ~8 mg/L for a mid-log culture of OD 2 at 750 nm. Under ideal growth conditions, this equates to approximately 17 mg Cpl-1 per gram dry weight, or 3.1 mg/ L/ day.

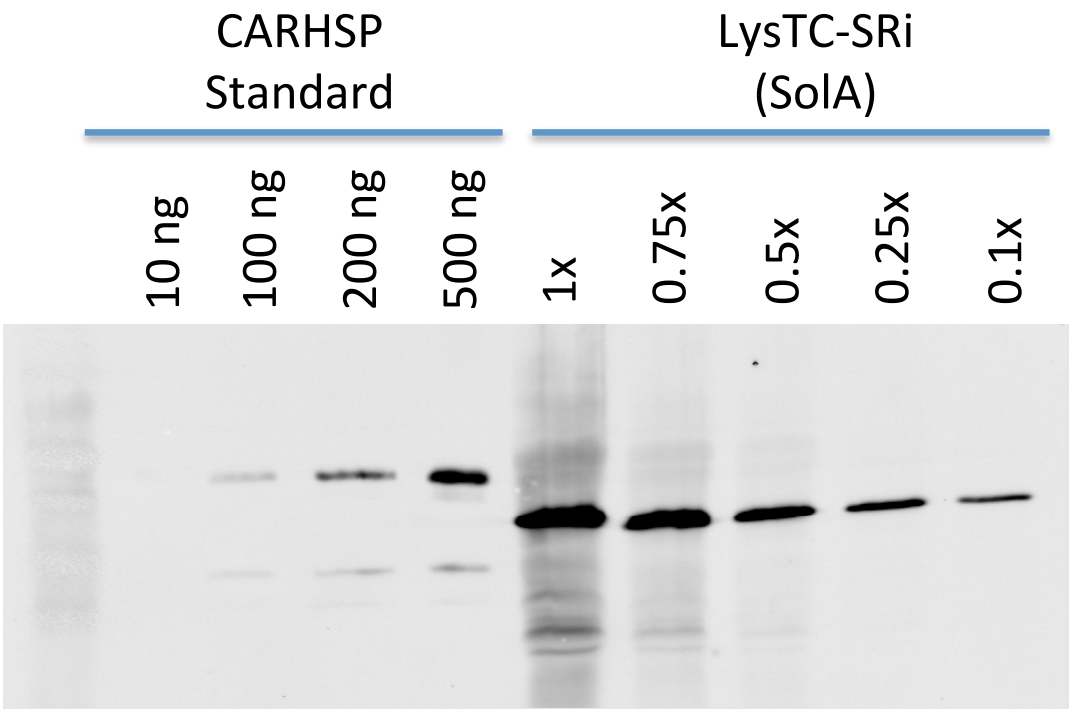


Figure 3.14 - Western blot analysis with anti-HA antibodies of a LysTC-SRi dilution series with commercial HA-tagged CARHSP as a reference

This side-by-side western blot analysis of LysTC-SRi and the commercially produced HA standard CARHSP allows for quantification of Cpl-1 accumulation. Cpl-1 accumulation in 1x LysTC-SRi is shown to be considerably higher than that of the highest concentration reference sample (500 ng).

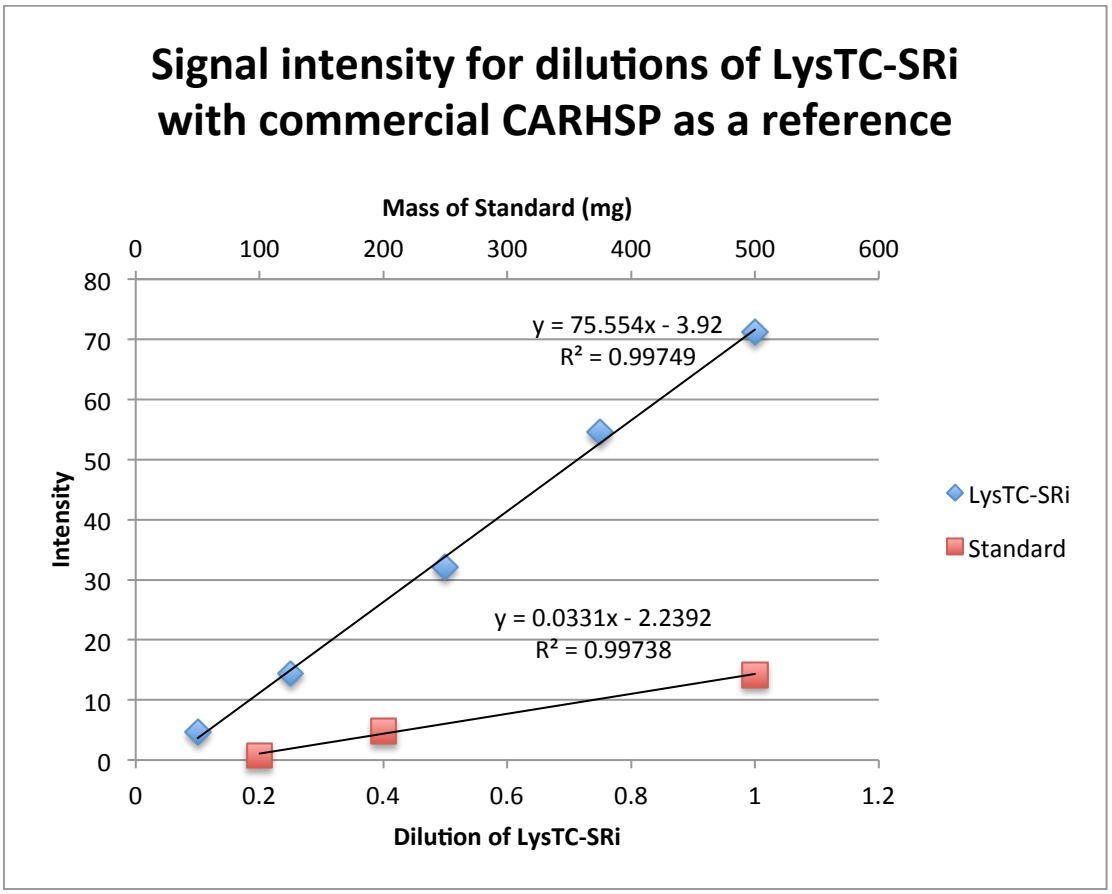


Chart 3.3 - Dilutions of LysTC-SRi with commercial CARHSP as a reference

From quantitative analysis of the western blot analysis shown in Figure 3.14 with the LiCor Odyssey system it can be seen that levels of both HA tagged proteins correlate very closely with their respective dilutions with high R^2 values in each case.

Table 3.2 - Quantitative analysis of Cpl-1 content of *LysTC-SRi*

LysTC-SRi dilutions were analysed for HA-tagged protein content using the LiCor Odyssey scanner. Intensity readings were converted into masses by extrapolation of the line of best fit for the HA-tagged standard, CARHSP, taking into account the slightly smaller mass of Cpl-1 and weighted by dilution factor. Mass of HA-tagged protein was then scaled to mg/L for a 10x cell culture solution for comparison to the total soluble protein measurements above.

LysTC-SRi Dilution	Intensity	ng/ 25 µl	mg/L 10x concentration	%TSP	Weighted %TSP	mg/L 1x concentration
1	71.21	2017	80.69	9.42	9.42	8.07
0.75	54.54	1559	62.38	7.28	9.70	8.32
0.5	32.1	943	37.73	4.40	8.80	7.55
0.25	14.42	458	18.30	2.14	8.54	7.32
0.1	4.57	187	7.48	0.87	8.73	7.48

Weighted Mean	9.04	7.75
Standard Deviation	0.50	0.42

3.2.3.3 ***Quantification of Cpl-1 is not verified by Coomassie stained SDS-PAGE***

The value seen for %TSP, though comparable to other proteins quantified in the Purton lab using the same method, is high in relation to %TSP values reported by other groups (see 1.3.3) (Specht *et al.*, 2010). To corroborate the data, an ultracentrifuged protein preparation from 3.2.3.1 was run on an SDS-PAGE alongside the commercial HA standard and the resulting gel stained with Coomassie and analysed using the LiCor Odyssey system as described above (Figure 3.15). From data presented in Figure 3.14 and Table 3.2, the 1000 ng HA standard band of Figure 3.15 should be approximately half the intensity of the Cpl-1 band from *LysTC-SRi* (assuming equivalent efficiency of Coomassie staining for the two proteins). It is not clear however, if a Cpl-1 band is visible at all for LysTC-SRi, let alone one of the predicted intensity. This casts doubt over the quantification figures presented in Table 3.2. It should be noted that although %TSP is widely used in the *C. reinhardtii* community, a universal protocol for its derivation is yet to emerge.

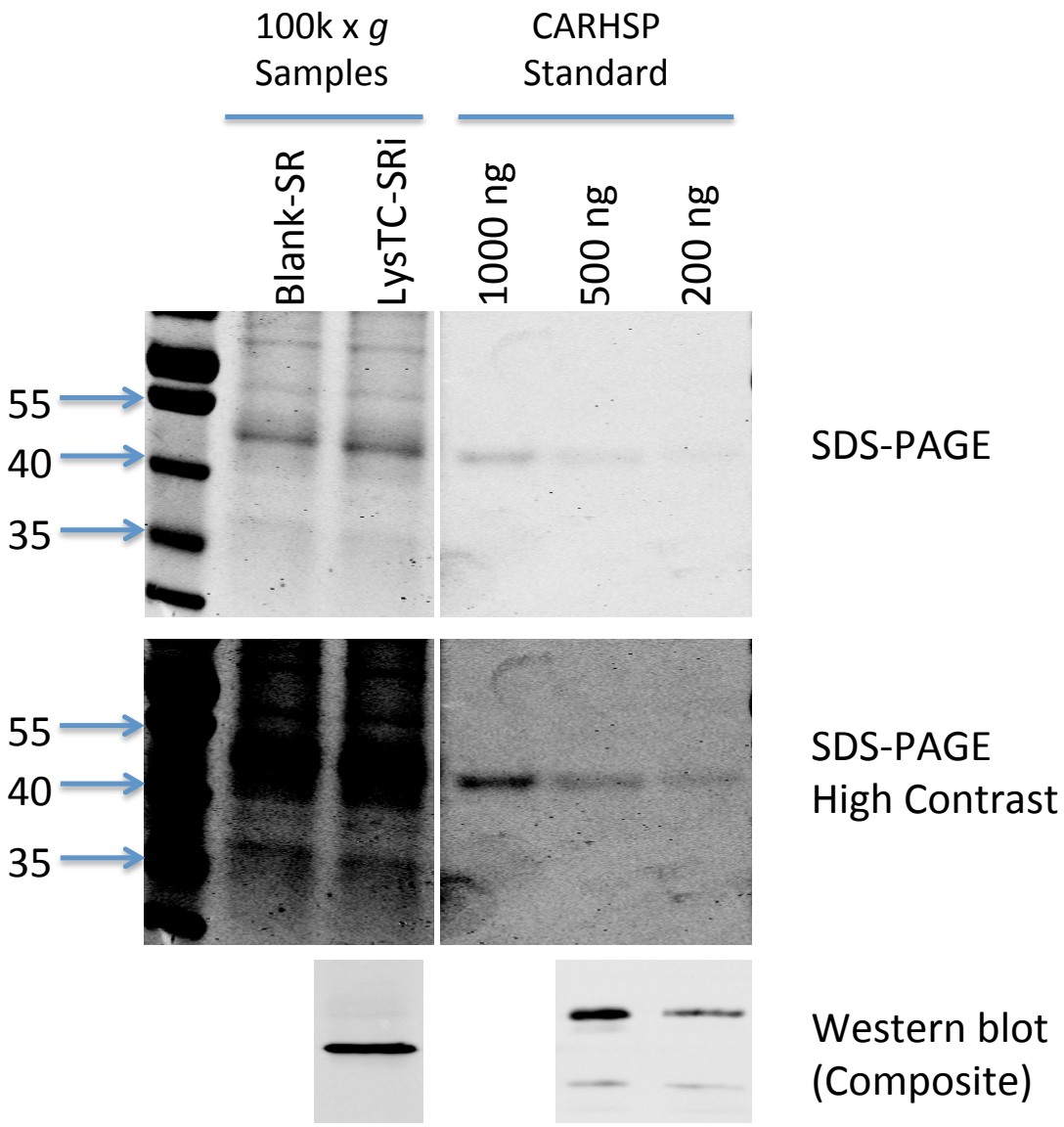


Figure 3.15 – Coomassie Brilliant Blue R stained SDS-PAGE of blank (Blank-SR) and *cpl*-1 expressing (LysTC-SRi) in relation to the CARHSP standard.

Top panel: Standard contrast SDS-PAGE stained with Coomassie Brilliant Blue R and visualised by the LiCor Odyssey system. Major bands are faintly visible for all three of the CARHSP standard lanes, however Cpl-1 is not.

Middle panel: High contrast image of the same SDS-PAGE. Here the CARHSP bands are clearly visible, however there is no indication of Cpl-1 in the predicted 38 – 40 kDa region.

Lower panel: Western blot analyses with anti- HA antibodies of LysTC-SRi and CARHSP at the same concentration as in the above SDS-PAGE. It is clear there is a discrepancy in band intensities between western blot analysis and Coomassie Brilliant Blue R stained gels.

3.2.4 Investigations into production and maintenance of non-denaturing protein preparations

As no lysins had previously been expressed in the *C. reinhardtii* chloroplast, a key step was demonstrating that the recombinant Cpl-1 produced was folding into an active state. Confirmation of expression and analysis of Cpl-1 accumulation were both conducted by western blot analysis, permitting a denaturing method of cell disruption using SDS. Purification and activity assays, however, would require the production of non-denatured protein extracts, thus necessitating a mechanical cell breakage protocol to be developed.

3.2.4.1 ***Non-denaturing protein preparations of LysBCi are produced by cell disruption***

As discussed in Chapter one, the *C. reinhardtii* cell wall is protein based and thus is readily solubilised by boiling in SDS. In order to produce non-denatured protein extracts however, a mechanical means of cell breakage is required. Various strategies were investigated to break the cell walled *cpl-1* expressing line LysBCi, including sonication (using either a sonication bath or a probe) and freeze-thawing. However, a pressured based lysis approach using the OneShot cell disruption system was found to be most effective. Late log cultures were concentrated by a factor of 100x relative to initial culture volume and disrupted at 30k PSI. The protein extract produced was highly dense and took several hours of low speed (10k x *g*) centrifugation to produce suspension-free extracts. During this process Cpl-1 was seen to localise to the pellet as opposed to staying in solution as reported in other systems (Loeffler *et al.*, 2003; Oey *et al.*, 2009b). At this point it was unclear whether this was due to the production of insoluble protein aggregates or the binding of otherwise soluble Cpl-1 to insoluble cellular components.

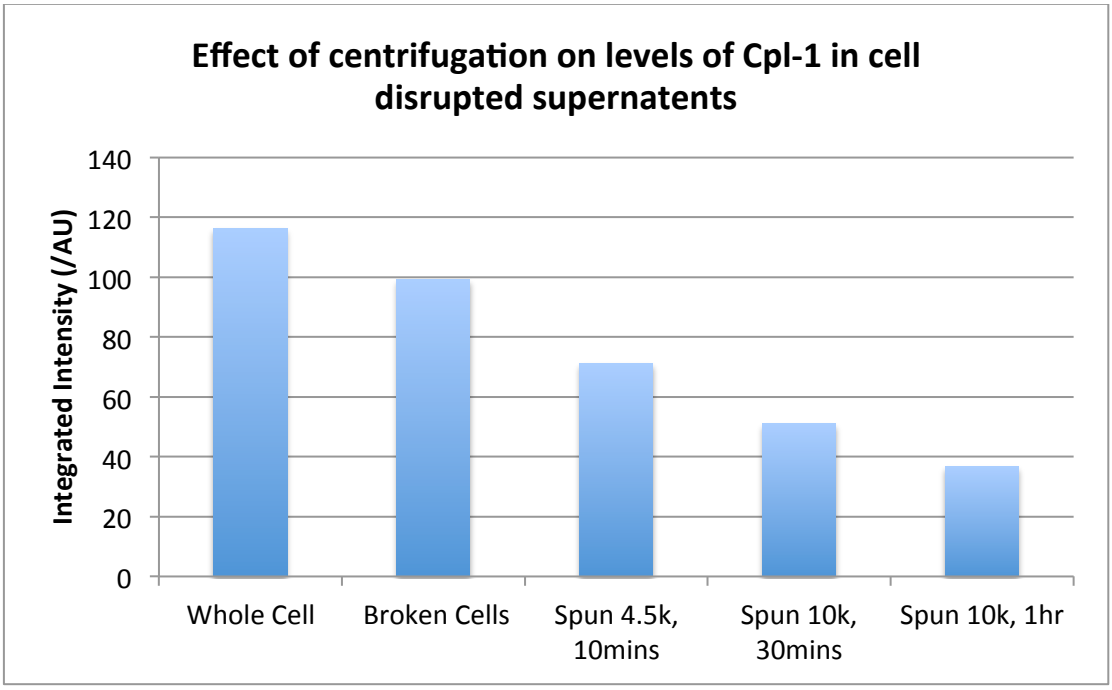


Chart 3.4 – Progressive loss of Cpl-1 from cell disrupted supernatants during low speed centrifugation

Quantitative analysis of Cpl-1 in the supernatant of broken LysBCi samples at various time points during a low speed centrifugation shows the progressive loss of Cpl-1 to the pellet as the cell debris is cleared.

The association of Cpl-1 with the cell debris suspension severely hindered both purification (where columns were quickly clogged), and absorbance based activity assays (where particulate light scattering masked results). Investigations into cell disruption in the presence of various detergents and co-factors were conducted. These included the non-ionic detergents Tween-20, Triton X-100, and CHAPS to disrupt non-specific interactions, and choline chloride to interrupt any targeted interactions with the Cpl-1 choline binding domains (data not shown). When none of these resulted in solubilised Cpl-1, a dilution-based approach was considered following observations made when optimising protein extraction in *E. coli* for an unrelated project. Protein extracts were prepared by cell disruption at 5, 10, and 100x concentrations relative to the cell culture. Broken cells were then centrifuged until a clear supernatant was achieved, and the concentration of Cpl-1 in both supernatant and pellet fractions analysed by western blot analysis. Quantitative analysis was conducted using the LiCor Odyssey system as previously described. From the data presented in Figure 3.16 and Chart 3.5 it is clear that high cell densities cause Cpl-1 to precipitate out of solution. Owing to the volume-limited nature of the cell disruption apparatus (the maximum volume of the chamber is 8 ml), such low concentrations prohibited production of protein extract at any kind of scale. To address this issue the experiment was repeated with samples disrupted at 100x cell culture concentration followed by dilution of broken samples directly before centrifugation. This was shown to be equally effective as diluting samples prior to disruption (data not shown).

It was also found that production of suspension free supernatant samples could be achieved more easily by freezing and thawing of the disrupted sample followed by centrifugation at low speed (4.5k x *g*, 10 minutes). Overnight incubation at 4 °C was also found to be effective, as was incubation on ice for 30 minutes.

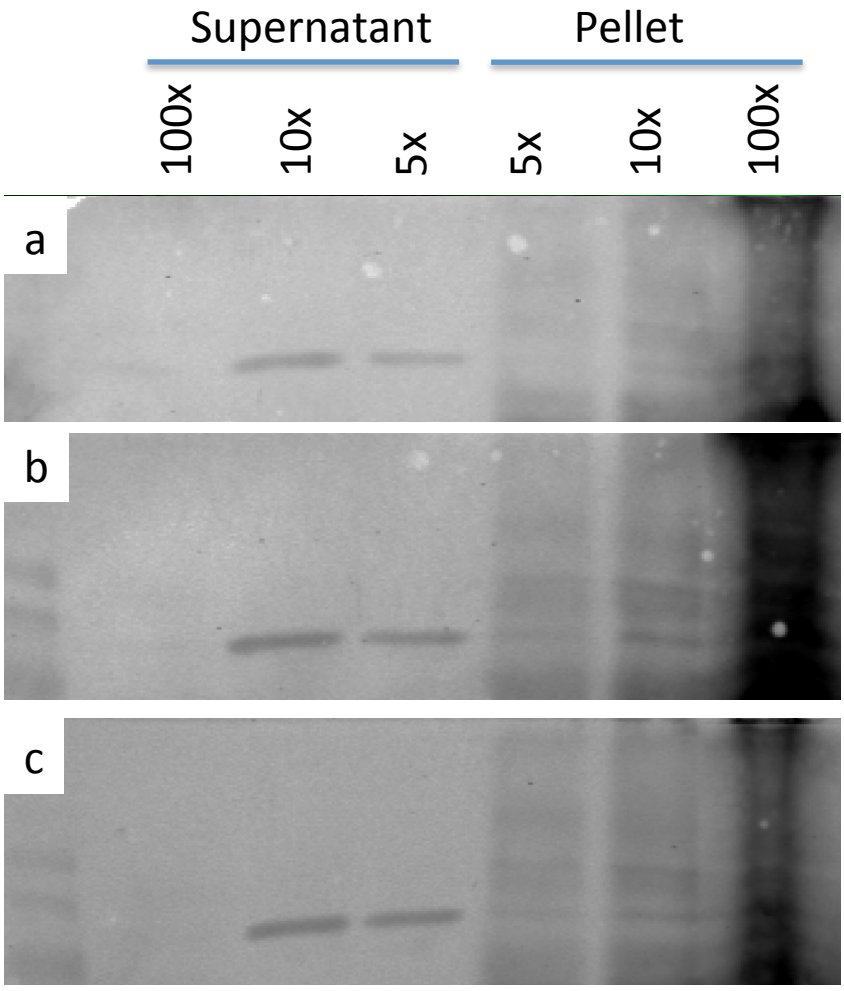


Figure 3.16 – Triplicate western blot analysis with anti-HA antibodies illustrating the effect of cell concentration on Cpl-1 supernatant retention

Pellet samples are obscured by the large amount of chlorophyll and other cell debris necessarily loaded with the pellet, however the supernatant samples clearly shown the presence of more Cpl-1 at lower cell breakage concentrations.

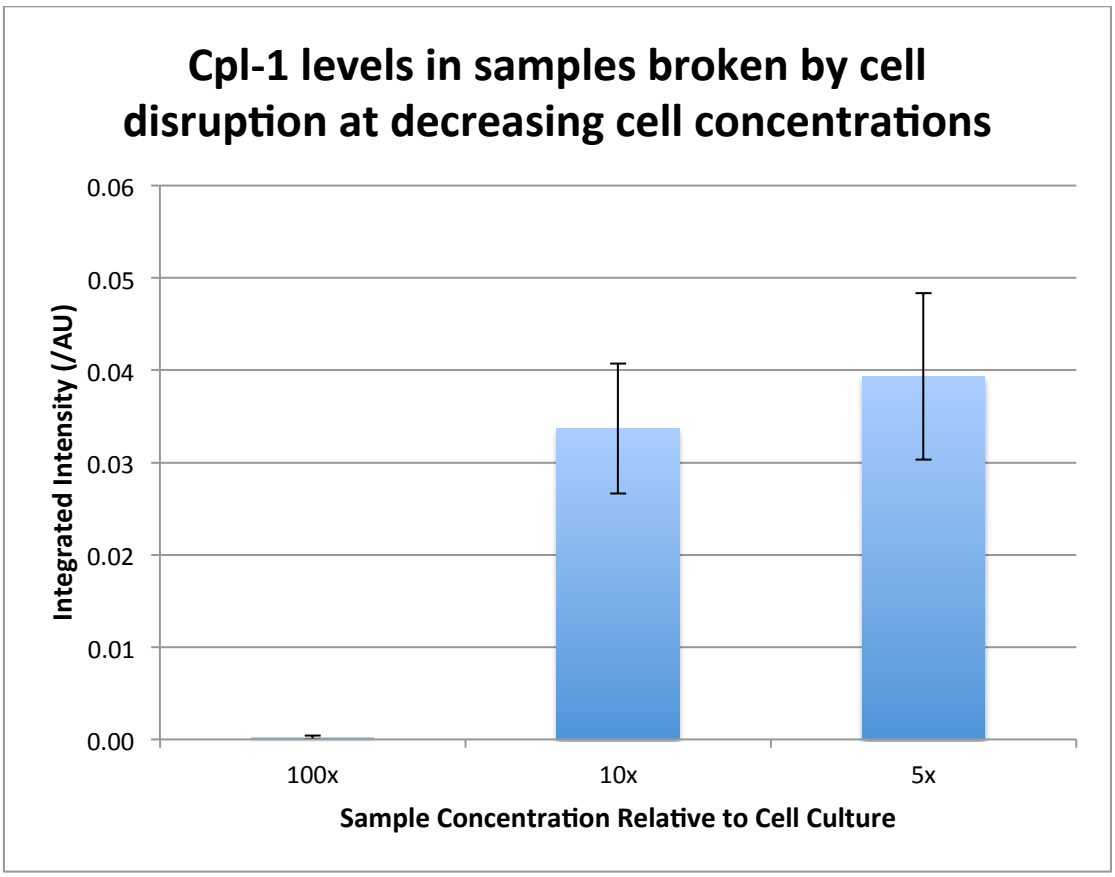


Chart 3.5 – Quantitative analysis of the effect of cell concentration on Cpl-1 supernatant retention

Quantitative analysis of Figure 3.16 shows the extent to which Cpl-1 is lost to the pellet on centrifugation. Levels of Cpl-1 in the 100x sample should be 10 fold that seen in the 10x sample, which in turn should be twice those of the 5x sample. The trend observed is contrary to this, suggesting that in the broken cell environment Cpl-1 solubility is inversely dependent on sample concentration.

3.2.4.2 ***Investigations into production of non-denaturing protein preparations from the cell wall deficient line LysTCi***

Expression of *cpl-1* in the cell wall-deficient background of TN72 (as opposed to the cell walled bst-same1) allows for techniques other than pressure cell disruption to be used to produce non-denatured protein preparations. To assess the relative efficacy of the different cell breakage techniques available, and to ascertain if Cpl-1 was still being lost to the pellet in the cell wall-less environment, a late log LysTCi culture was split and subjected to a panel of cell breakage techniques (listed below). The culture was equalised and resuspended in PBS (unless stated otherwise) to a concentration of 10x or 100x relative to culture volume.

1. Freeze/ thaw at -196 °C/ 30 °C (3 rounds)
2. Freeze/ thaw at -20 °C/ 30 °C (1 round)
3. Cell disruption at 10k PSI
4. Addition of 1 % Triton-X 100 (30 minute incubation at room temperature)
5. 1 minute vortex (max speed)
6. No treatment (30 minute incubation at room temperature)
7. Resuspension in ddH_2O (30 minute incubation at room temperature)

Following the above treatments samples were centrifuged, and supernatants stored at -20 °C overnight, with the exception of sample 3, which was incubated at 4 °C overnight prior to centrifugation. Samples were then analysed by western blot and quantitative data collected using the LiCor Odyssey system as described above (2.4.3).

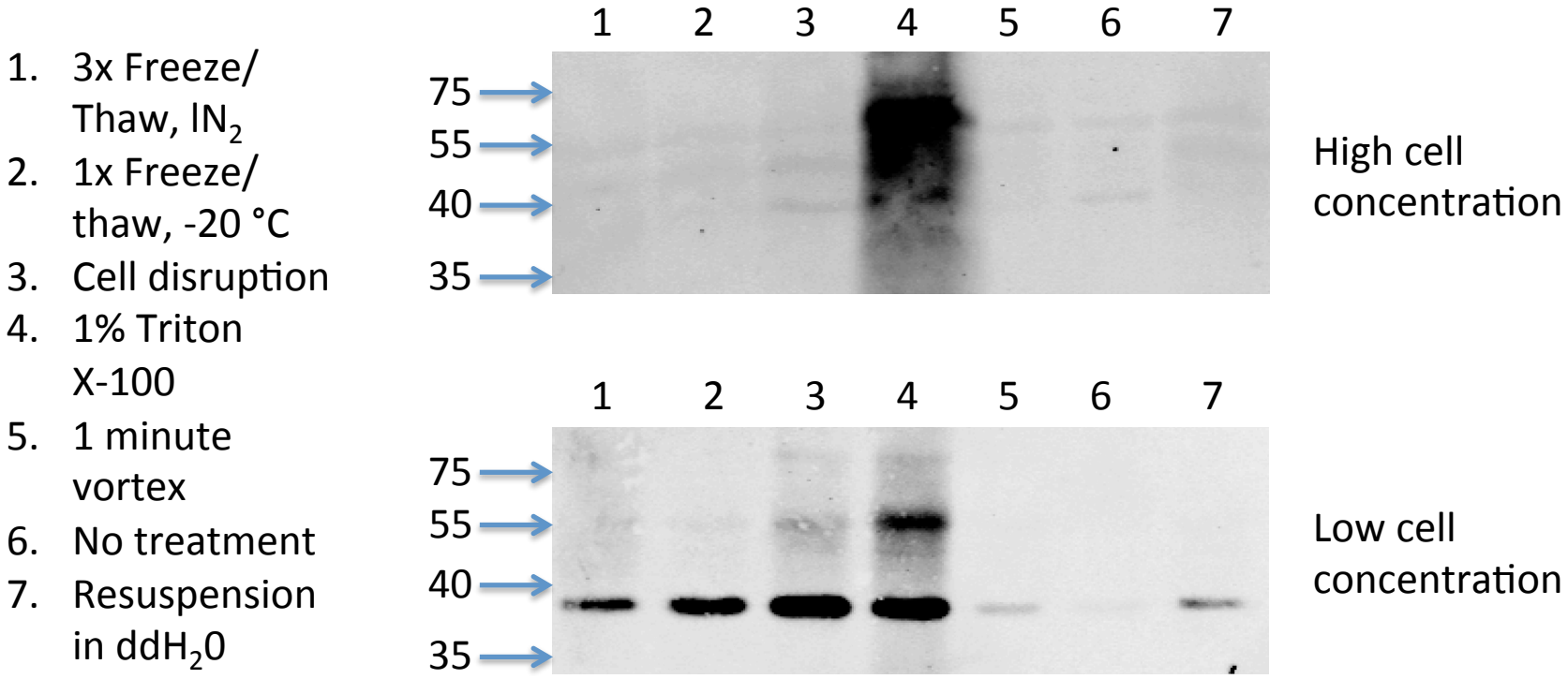


Figure 3.17 – Retention of Cpl-1 in the supernatant following various cell breakage techniques at 10 and 100x culture concentration (LysTCi expression)

Top panel: Cell breakage and centrifugation at high cell concentration (100x cell culture, ~1.84*10^9 cells/ml) continues to give extremely poor yields of Cpl-1 for all cell breakage techniques.

Lower panel: Cell breakage and centrifugation at low cell concentration (10x cell culture, ~1.84*10^8 cells/ml) shows the presence of Cpl-1 for all samples including the untreated control.

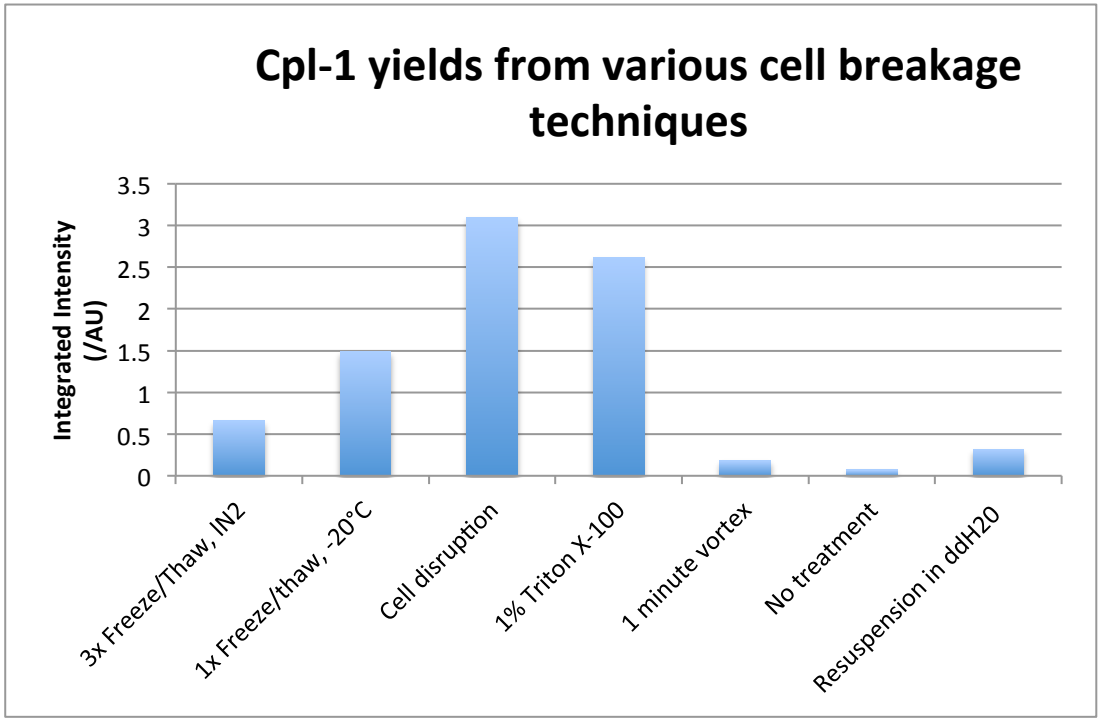


Chart 3.6 – Graphical representation of retention of Cpl-1 in supernatant following various cell breakage techniques at 10x culture concentration as seen in Figure 3.17

Cell disruption and use of detergent are shown to be the most effective means of breaking cells with the TN72 cell wall deficient phenotype.

It is clear from Figure 3.17 that issues surrounding solubility of Cpl-1 at high concentrations are not restricted to the cell walled transformant line LysBC, or the cell disruption technique. The loss of Cpl-1 from the supernatant is evident in all samples analysed at 100x cell culture concentration including chemical disruption with a mild detergent. At 10x concentration however the presence of Cpl-1 is observed in all cases. From Chart 3.6 it appears that cell disruption gives the highest yield of Cpl-1, higher even that Triton-X 100 treatment, despite microscopic analysis of Triton-X 100 treated cells showing complete cell breakage (data not shown). A single freeze/thaw overnight at -20 °C is shown to be significantly more effective than three brief rounds at -196 °C, however it still does not yield as much Cpl-1 as cell disruption. It is possible that multiple rounds of -20 °C freezing and thawing, or a single -20 °C treatment followed by multiple rounds at -196 °C may be of comparative efficacy to cell disruption. Such a procedure would be preferable as it is less labour intensive and also allows for samples to remain sterile during the cell breakage process. Vortexing and resuspension in ddH_2O, though non-viable as cell disruption techniques, are interesting as an illustration of the fragility of TN72 based lines in general.

*3.2.4.3 **Investigations into Cpl-1 stability in a* C. reinhardtii *crude cell extract environment***

Once Cpl-1 containing protein preparations could be reliably produced, investigations were conducted into the stability of Cpl-1 in a crude cell extract. It was known that Cpl-1 was highly stable in the *S. pneumoniae* cytoplasm, and also in the unbroken *C. reinhardtii* chloroplast (C. Economou, unpublished work); however, it was unclear how stable Cpl-1 would be once it had been exposed to the alien environment of the eukaryotic cytoplasm. This is clearly important from a pharmaceutical prospective, especially when considering a topical *C. reinhardtii* based treatment where only crude extracts might be employed.

As a preliminary investigation into the level of stability shown by Cpl-1 in a broken cell environment, a cell extract was prepared as previously described for quantification of TSP (3.2.3.1), with the exclusion of the centrifugation step. Samples were treated with chloramphenicol to remove any possibility of continued protein synthesis, and incubated over a course of 93 hours at 4 °C, 25 °C,

and 37 °C. From these preliminary data (Figure 3.18, Chart 3.7) it is clear that Cpl-1 is relatively stable even when in a broken cell environment. It is also shown that degradation is highly temperature dependant. Almost no appreciable lost of product is seen for the 4 °C sample over the time course, whereas a steady decline is seen for 25 °C with very little protein present after 3 days, and 37 °C showing almost complete degradation in under 24 hours. These data suggest suitable levels of stability for clinical use, both in the context of pre-application cold storage, and short-term topical application. Use at physiological temperatures, for example by intravenous or intraperitoneal administration, is likely to be unaffected by the rapid temperature based degradation, as from previous studies it is clear that immune based clearance will be significantly faster (see 1.1.4.3).

3.2.5 Purification of Cpl-1 by ion exchange chromatography

The purification of *C. reinhardtii* derived Cpl-1 can be seen as important for a number of reasons. Primarily it is an important step in the long-term goal of this project - as a therapeutic agent it will almost certainly require some form of purification or at least enrichment of the active product. The demonstration of such a protocol is essential to the case of *C. reinhardtii* as a platform for next generation anti-bacterials. Purification is also important to ensure that enrichment of Cpl-1 correlates with enrichment of activity, and for more detailed enzymatic and antimicrobial assays than is possible with crude extracts.

The protocol for purification of Cpl-1 from *C. reinhardtii* was based on that used by Loeffler and colleagues for Cpl-1 produced in *E. coli*; namely ion exchange with choline as a specific elutant (Loeffler *et al.*, 2001). Various other methods of enrichment and purification were investigated, including ammonium sulphate precipitation, gel exclusion chromatography, and HA- affinity chromatography; however, for reasons of resolution, expense, and time constraints, ion exchange was chosen for a focused study.

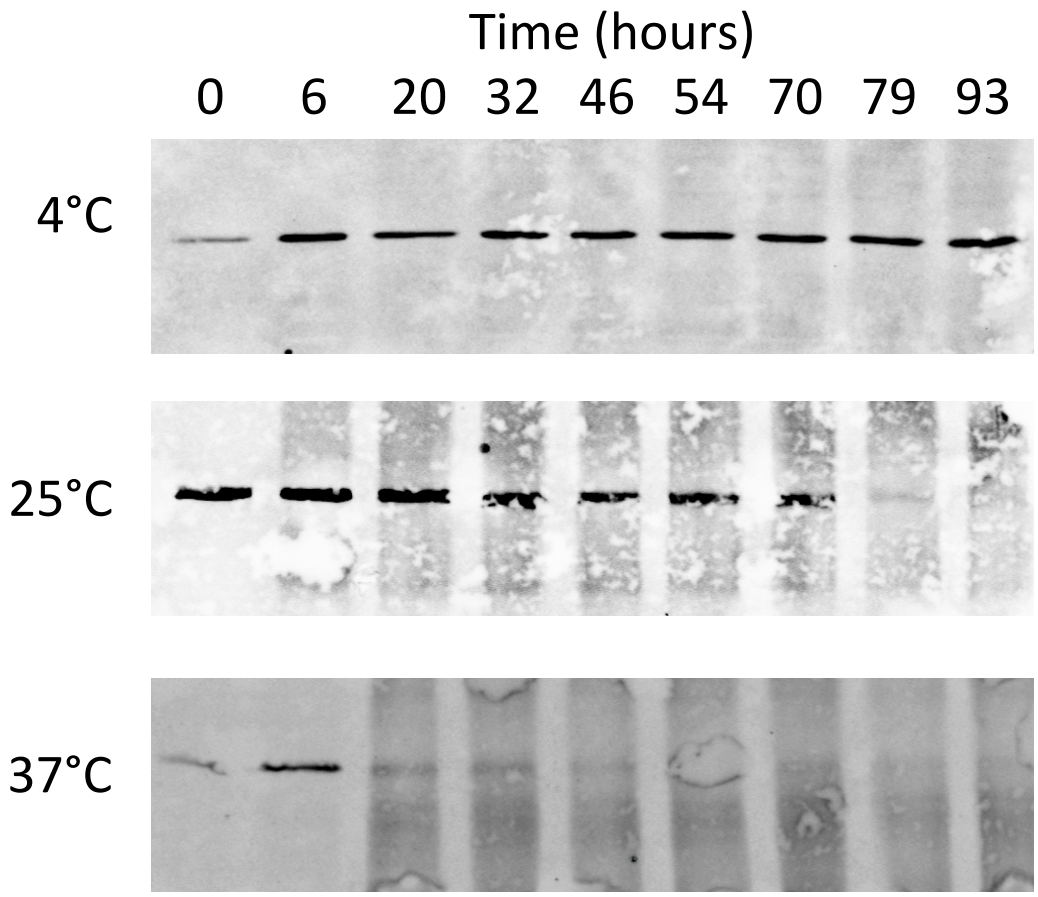
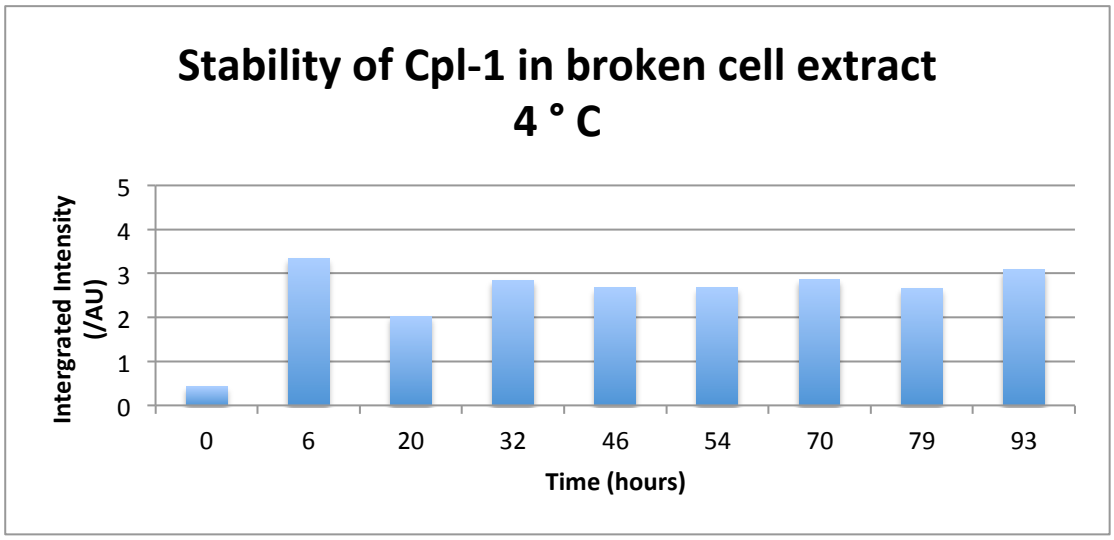
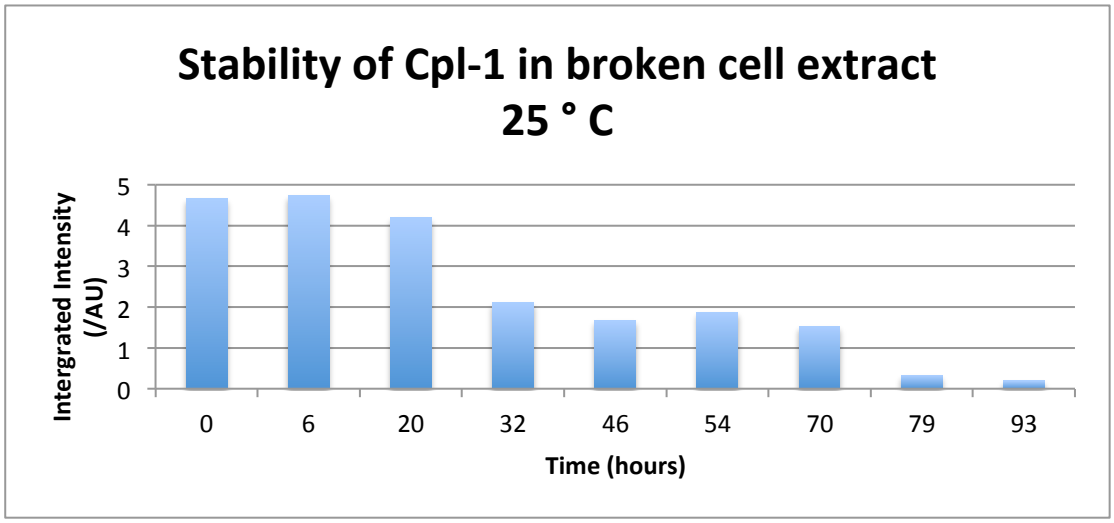


Figure 3.18 – Western blot analysis with anti-HA antibodies showing preliminary stability analysis of Cpl-1 in a crude cell extract at 4, 25, and 37 °C

Extracts were equalised to 10x cell culture concentration and broken by cell disruption at 10k PSI. Un- centrifuged 1 ml samples were then incubated in the dark at the above temperatures and 50 µl samples taken periodically and stored at -20 °C. Samples at 4 °C show the least Cpl-1 degradation, with 25 and 37 °C showing progressively more. The cause of the low levels of Cpl-1 in t=0 for 4 and 25 °C is not known, but is considered an artefact.





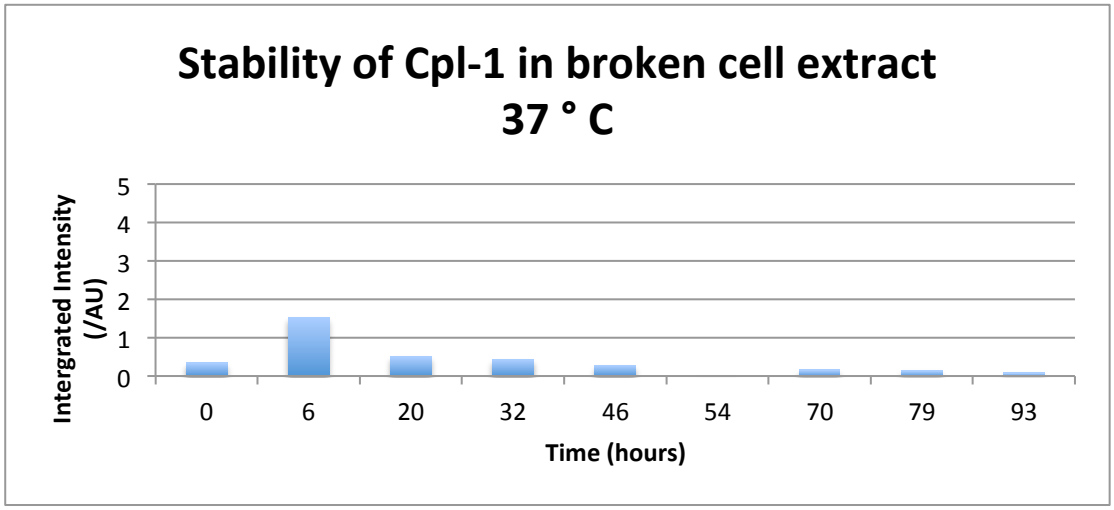


Chart 3.7 – Graphical representation of stability of Cpl-1 in a crude cell extract at multiple temperatures from Figure 3.18

With the exceptions of t=0 for 4 °C and 37 °C these data, though preliminary, show high stability at 4 °C, progressive degradation after 24 hours at 25 °C, and rapid degradation from the start at 37 °C.

3.2.5.1 ***Binding of Cpl-1 to the column was shown to be compromised for early cell extracts***

Initial attempts at purification were conducted prior to the resolution of the solubility issues described in section 3.2.4. Columns were loaded with low speed centrifuged samples containing high levels of particulate matter. Although this approach was successful in purifying small amounts of Cpl-1, columns became rapidly saturated resulting in large portions of Cpl-1 being lost in flow-through fractions (Figure 3.19, Chart 3.8). Once the issue of solubility had been resolved however, analysis of individual (Figure 3.20) and pooled flow through and wash fractions (Figure 3.21) displayed no evidence of further loss of Cpl-1 from the column.

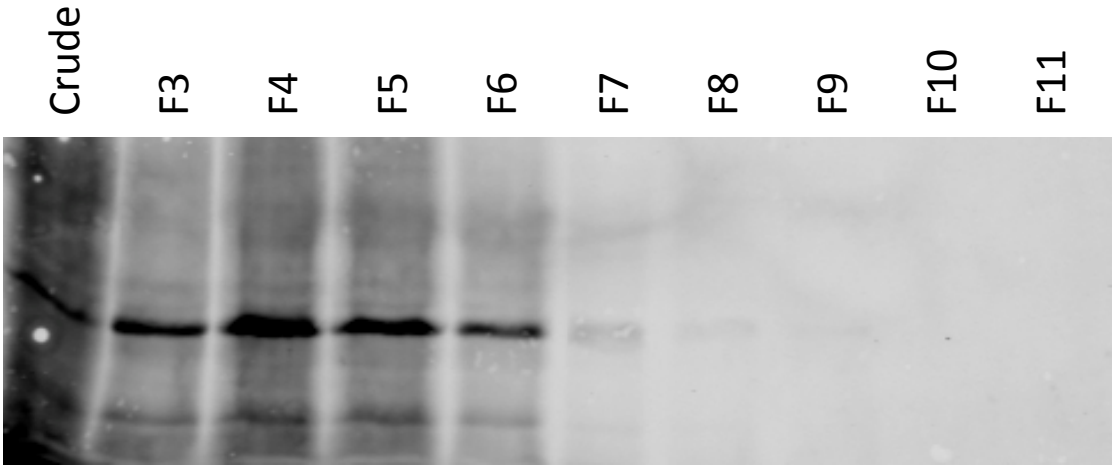


Figure 3.19 – Western blot displaying loss of Cpl-1 in flow-through fractions during DEAE cellulose purification of LysBCi before resolution of solubility issues.

Fractions F3-F6 show Cpl-1 leaving the column in the flow through indicating insufficient binding capacity of the DEAE matrix. After the initial flow through, no more Cpl-1 was lost in the wash stages.

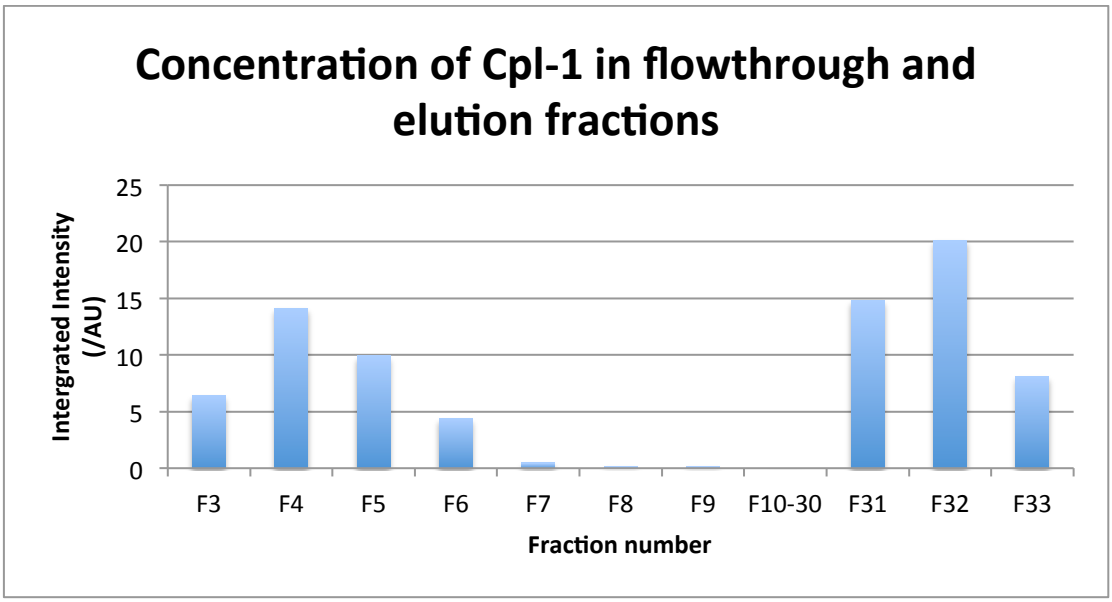


Chart 3.8 – Graphical illustration of loss of Cpl-1 in flow-through fractions during DEAE cellulose purification

Although Cpl-1 is binding to the column as evidenced by its presence in the elution fractions (F31-F33), it is clear that a large portion of the recombinant protein loaded to the column is being lost in the flow through (F3-F6). Due to the small amounts of Cpl-1 involved, this is thought to be due to blocking of the column by cell debris as opposed to saturation by Cpl-1.

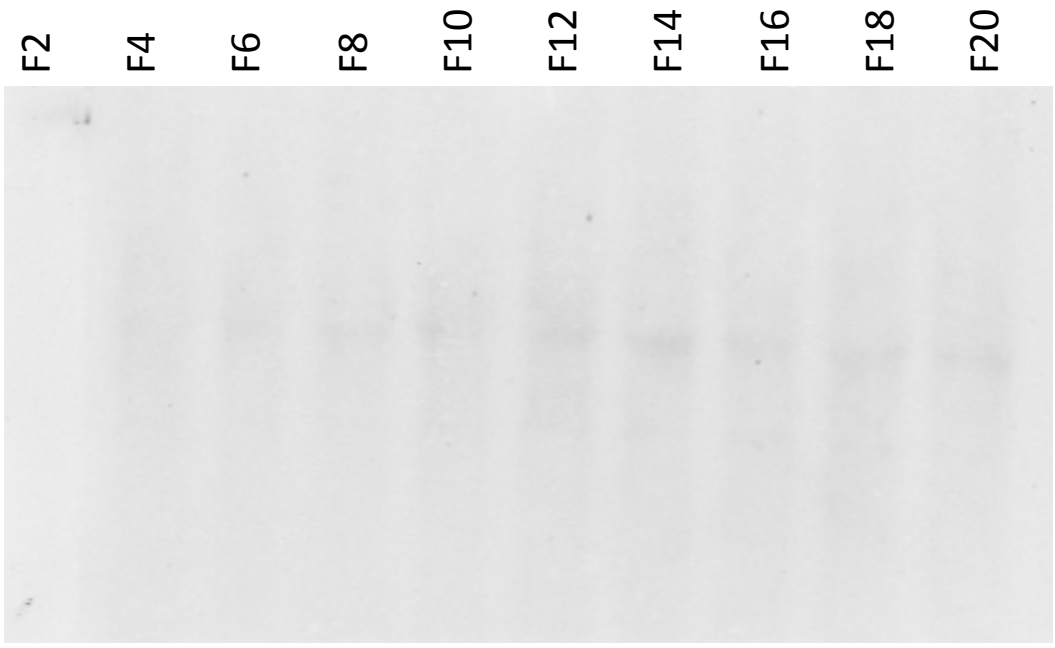


Figure 3.20 – Flow through fractions after resolution of solubility issues showing no loss of Cpl-1

With the resolution of both the issue of Cpl-1 solubility, and persistence of particulate matter in cell disruption supernatants, flow through fractions show no presence of Cpl-1.

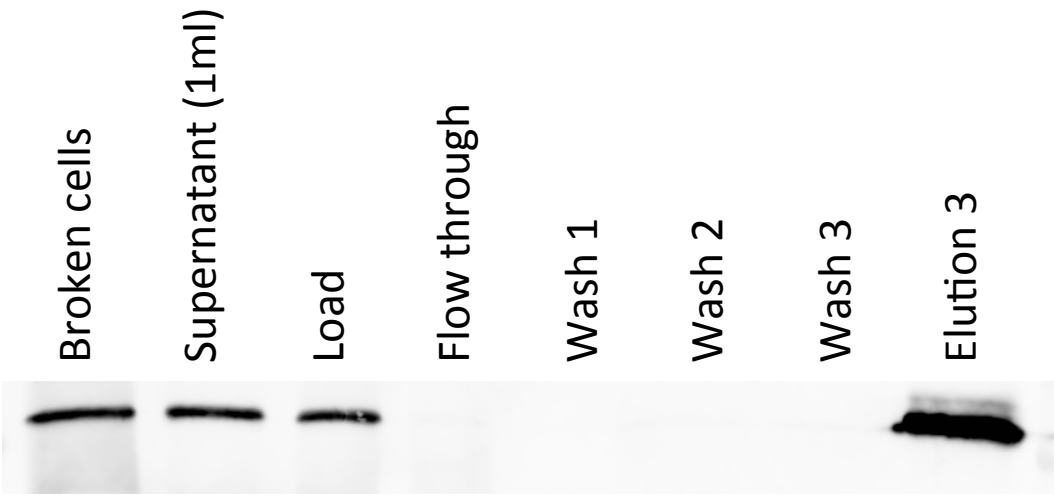


Figure 3.21 – Pooled flow-through and wash fractions showing no loss of Cpl-1 from the column

The absence of Cpl-1 from flow though and wash fractions is also evident when fractions are pooled.

*3.2.5.2 **Specific elution of Cpl-1 with choline chloride***

Elution of Cpl-1 was achieved using choline to specifically elute Cpl-1 by occupation of the choline binding sites in the C-terminal cell wall binding domain. Samples were collected and elution fractions analysed for the presence of Cpl-1 both by Bradford assay and anti-HA dot-blot (Figure 3.23). To assess the efficacy of removal of Cpl-1 from the column by choline addition, the column was subsequently washed with an excess of choline followed by a high salt (2 M NaCl) wash and the resulting fractions analysed by western blot. The absence of detectable protein following a high salt wash implies that all bound Cpl-1 has been removed from the column.

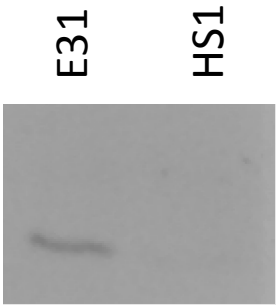


Figure 3.22 – Western blot showing levels of Cpl-1 in the final elution wash and high salt wash

Low levels of Cpl-1 are observed for the final elution wash, E31, consisting of two column volumes of elution buffer. The high salt wash (two column volumes 2 M NaCl in phosphate buffer) shows no Cpl-1 suggesting complete elution of Cpl-1 from the column.

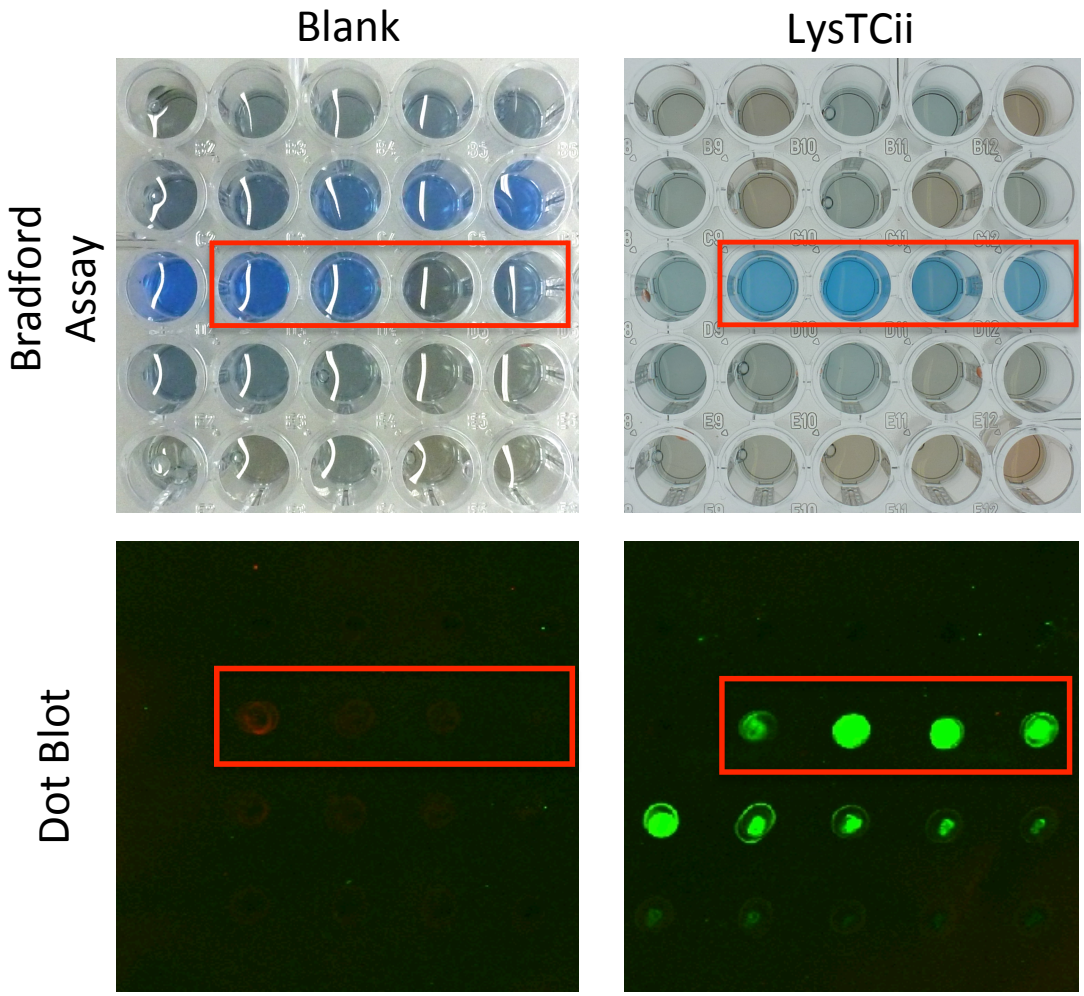


Figure 3.23 – Bradford assay identifying protein containing elution fractions, and anti-HA dot-blot analysis confirming the presence of Cpl-1

Initial identification of protein containing elution fraction was conducted by the Bradford assay, with an anti-HA dot-blot to confirm the presence of Cpl-1. In both cases the red box indicates fractions to be pooled and taken forward.

Top panel: Bradford assay to identify protein-containing fractions as indicated by a blue colour.

Lower panel: Anti-HA dot-blot to confirm presence of Cpl-1.

3.2.5.3 Purity analysis of enriched samples by SDS-PAGE

Cpl-1 containing fractions were then pooled and concentrated 10 fold by centrifugal concentration and assessed for purity by SDS-PAGE (Figure 3.24) and quantitative analysis using the LiCor Odyssey system (Chart 3.9). These data show Cpl-1 to be present at a level of approximately 62 % by mass of total protein, assuming equivalent binding of Coomassie Brilliant Blue R to the various proteins in the sample. This represents a significant enrichment of Cpl-1, though not necessarily of sufficient magnitude to be classed as purification. These data have demonstrated, however, that *C. reinhardtii* derived Cpl-1 can be selectively bound and eluted from a DEAE cellulose column. The overall purity of Cpl-1 could doubtless be increased with further optimisation of the purification protocol or the addition of a polishing step such as affinity chromatography. A collection of later fractions was also pooled, concentrated, and analysed in the same manner. This sample gave a Cpl-1 level of ~20 %, suggesting that not only yield but also relative purity decreases in the later elution fractions (data not shown).

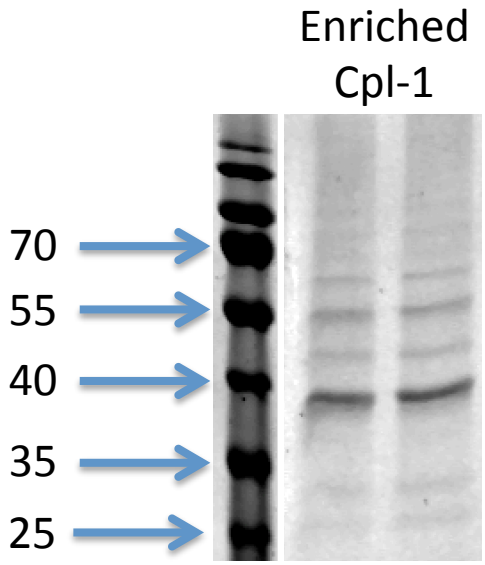


Figure 3.24 – Coomassie stained SDS-PAGE of concentrated pooled elution fractions (in duplicate)

Cpl-1 is shown to be the major protein present in the sample; however, several other contaminating proteins are also seen, most notably the RuBisCO large subunit at 56 kDa.

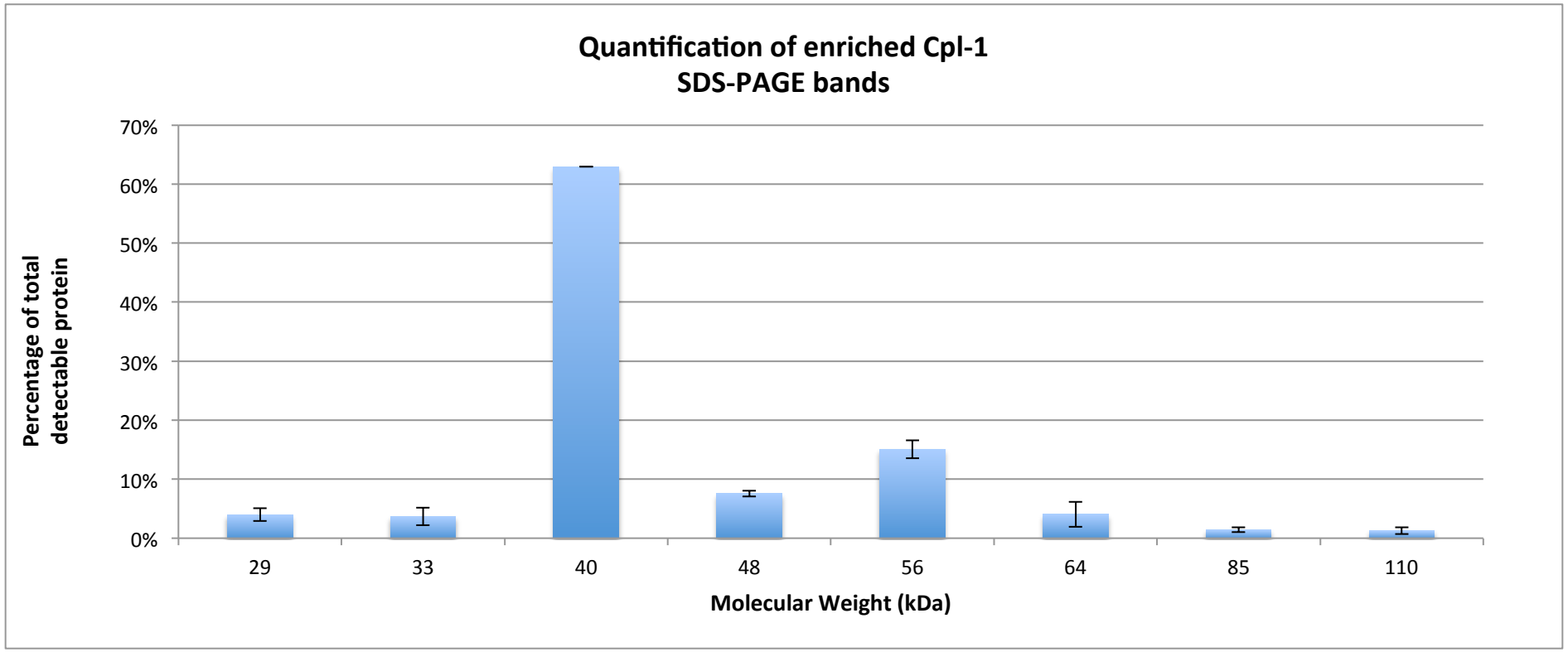


Chart 3.9 – Graphical representation of the relative levels of different proteins in an enriched Cpl-1 preparation.

Analysis of the data in Figure 3.24 by the LiCor Odyssey systems shows Cpl-1 to make up approximately 62 % of all protein in the sample by mass. With the exception of RbcL all other proteins are present at <10 % of total protein.

3.2.5.4 ***Dialysis for restoration of activity***

An issue reported by Loeffler and colleagues (Loeffler *et al.*, 2001) in regard to this method of Cpl-1 purification is that by using choline as an elutant, the enzyme's active sites are blocked, and thus all activity removed. To restore enzymatic activity the bound choline was removed by dialysis against enzyme buffer. For *C. reinhardtii* derived Cpl-1, pooled elution fractions were concentrated to a final volume of 3 ml and dialysed overnight at 4 °C. To ensure Cpl-1 was not inadvertently lost during the dialysis process, samples were taken before and after dialysis and investigated by western blot analysis (Figure 3.25). Although there is some loss of Cpl-1 in the dialysed sample relative to the pre-dialysis level (22 % decrease in this single experiment), this is not seen as significant enough to prohibit the use of choline as an elutant.

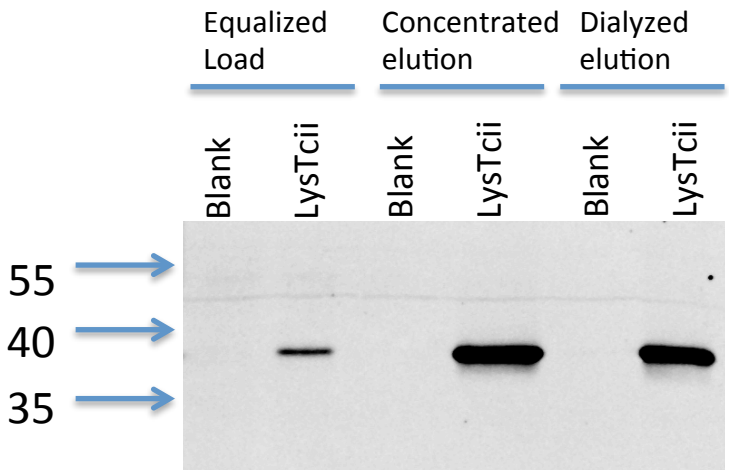


Figure 3.25 – Western blot analysis with anti-HA antibodies showing levels of Cpl-1 loaded onto the column, and before and after dialysis

This western blot analysis shows a small (~22 %) decrease in Cpl-1 after the dialysis process, however the post dialysis sample is shown to contain significantly more Cpl-1 than the initial load. Blank sample lanes pertain to TN72 transformed with the empty pSRSap1 vector, and prepared and purified in the same manner.

3.2.6 Demonstration of Cpl-1 activity against *S. pneumoniae*

The activity of *C. reinhardtii* derived Cpl-1 is central to this project, not just because it is critical to the use of *C. reinhardtii* for the production of this protein antibiotic, but also as a proof of concept for the production of active therapeutics in general using this platform.

3.2.6.1 ***Initial activity assays conducted on solid medium show endogenous antimicrobial activity of crude extracts***

For simplicity, early attempts to show *C. reinhardtii* derived Cpl-1 activity were conducted on solid media by spotting crude extract onto a bacterial lawn. As shown in Figure 3.26 such attempts were confounded by a previously unknown endogenous antibacterial activity in *C. reinhardtii*. It was hypothesised that such activity may be related to the nature of the cell wall defect seen in TN72, but as Figure 3.26 shows, the contaminating activity is also seen in the cell-walled wild type. Interestingly it is seen to a far lesser degree in a separate cell-wall deficient mutant cw-10*;* however, this was not investigated further due to time constraints.

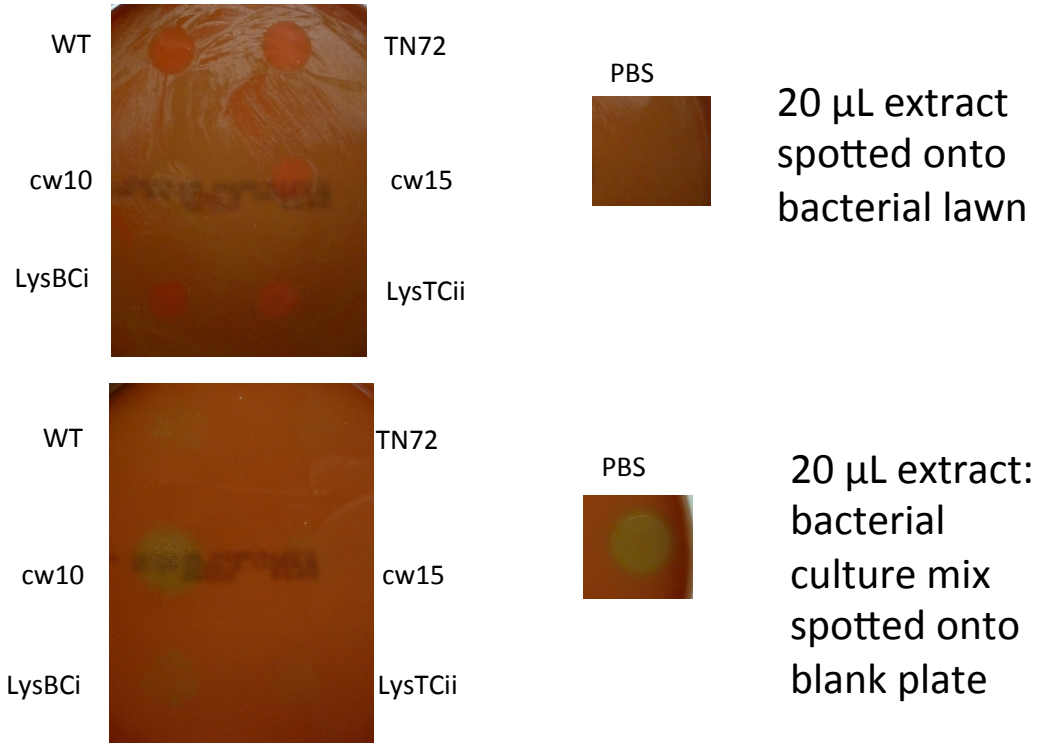


Figure 3.26 – Solid media activity assays for two lysin producing, and four non lysin producing *C. reinhardtii* lines

Top panel: Direct spot assays of 20 µl equalized protein extract onto a *S. pneumoniae* lawn. All extracts with the exception of cw10 show inhibition of *S. pneumoniae* growth.

Lower panel: Incubation assay of equalized protein extract with *S. pneumoniae* culture prior to spotting of 20 µl onto blank agar plate. As with direct spotting, all extracts show inhibition with the exception of cw10.

The endogenous activity observed prevents any conclusions from being drawn as to the activity of Cpl-1.

3.2.6.2 ***Clearance assays in liquid medium show Cpl-1 specific activity***

Turbidity based clearance assays, where activity is measured as a function of optical density of a bacterial suspension, were attempted at an early stage of this project prior to resolution of the solubility issue. These, however, gave inconclusive results as any drop in optical density was masked by high levels of particulate matter in suspension (data not shown).

Experimental work conducted in the Purton lab (L. Stoffels, unpublished work) suggested that though the endogenous activity was bactericidal, it did not in fact lyse the bacterial cells, and thus would not interfere with an optical density based clearance assay. Assays were conducted on both live and heat killed *S. pneumoniae* challenged with crude *E. coli* derived Cpl-1 (Figure 3.27), crude *C. reinhardtii* Cpl-1, and Cpl-1 enriched as described in 3.2.5 (Figure 3.28). Despite the decrease seen in blank samples (a result of known issues with *S. pneumoniae* autolysis), samples containing Cpl-1 show significantly larger decreases thereby confirming that Cpl-1 is active in each of the samples assayed.

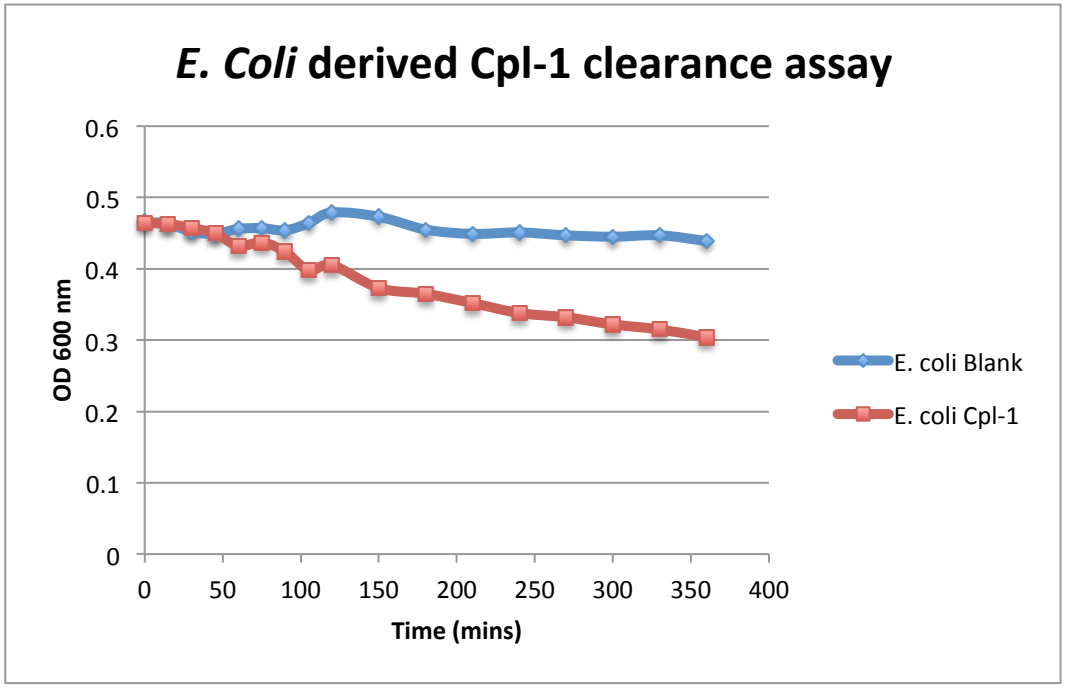


Figure 3.27 – Optical density based clearance activity assay for *E. coli* derived Cpl-1

Preliminary investigation into lytic action of Cpl-1 against *S. pneumoniae* shows a marked decrease in OD 600 nm for *E. coli* extracts containing Cpl-1 relative to a pASap1.empty negative control (*E. coli* Blank).

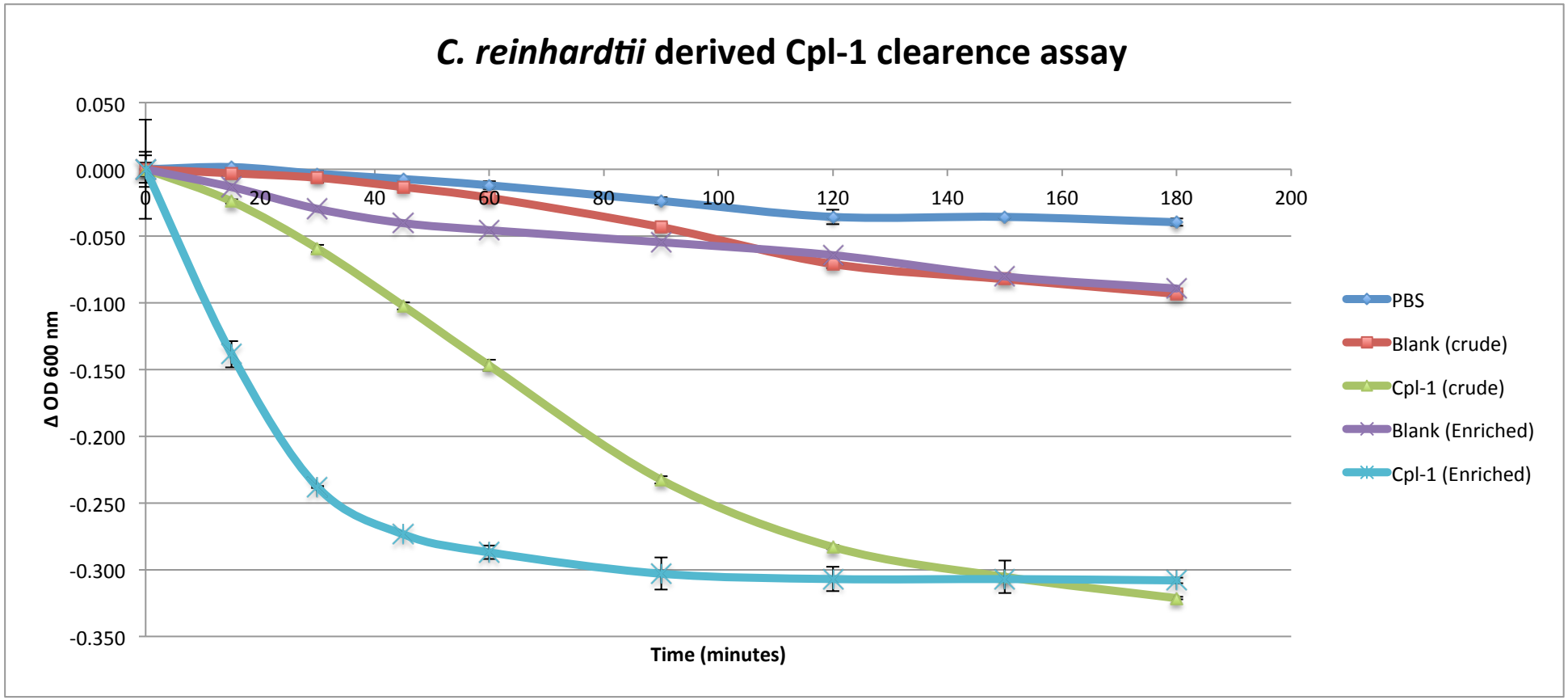


Figure 3.28 - Optical density based clearance activity assay for *C. reinhardtii* derived Cpl-1 in crude and enriched extracts

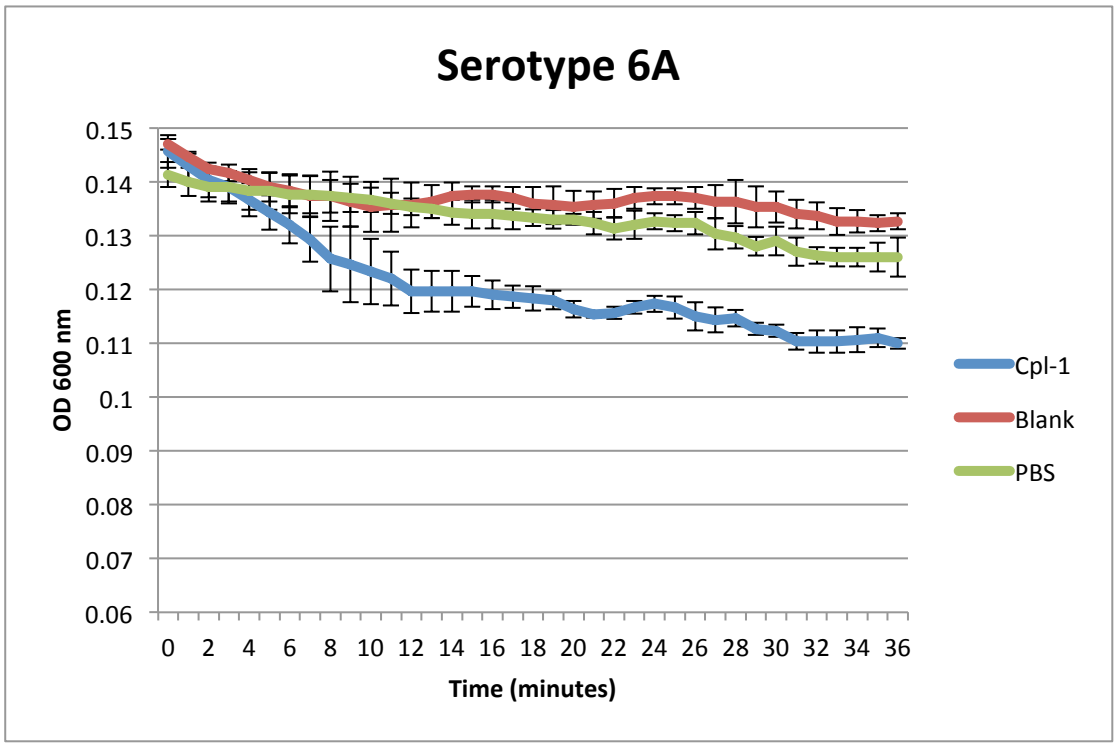
Both crude and enriched extract of the Cpl-1 expressing line LysTC-SRi show a significant clearance of *S. pneumoniae* culture relative to their corresponding blank samples (TN72 transformed with pASap1.empty), and a PBS negative control. The dose dependent variance in activity between crude and enriched samples acts as further evidence for the activity of *C. reinhardtii* derived Cpl-1.

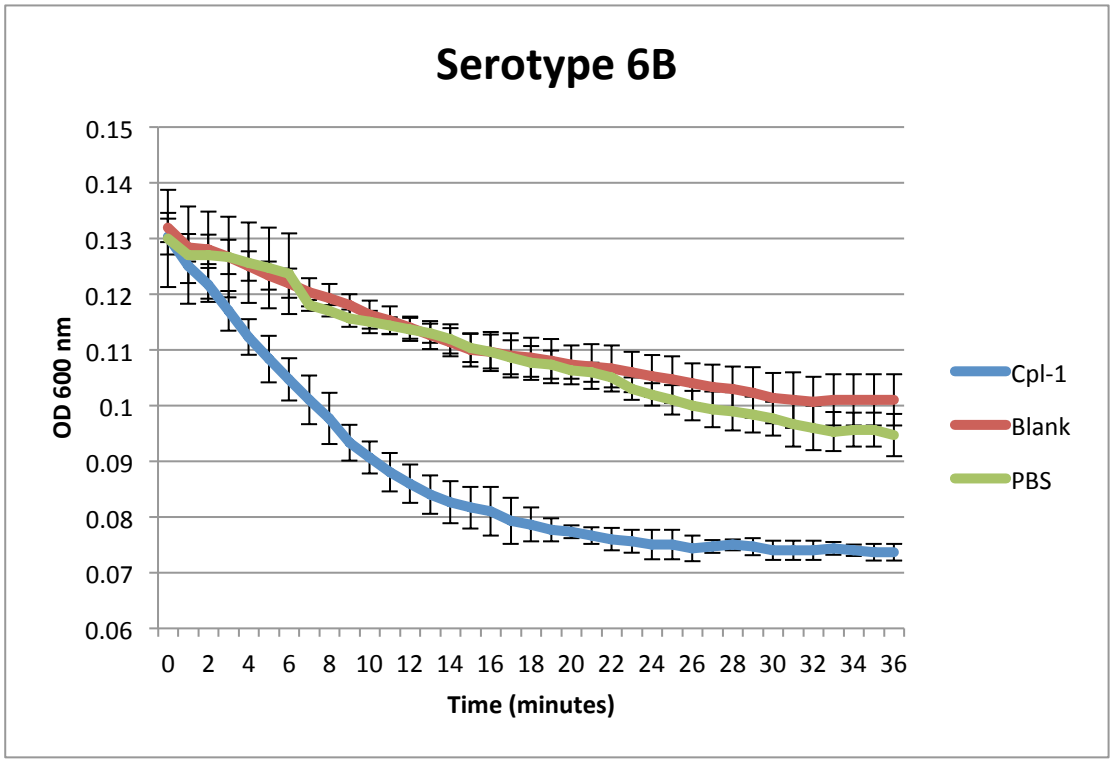
3.2.6.3 ***Demonstration of activity of C. reinhardtii derived Cpl-1 against clinical isolates***

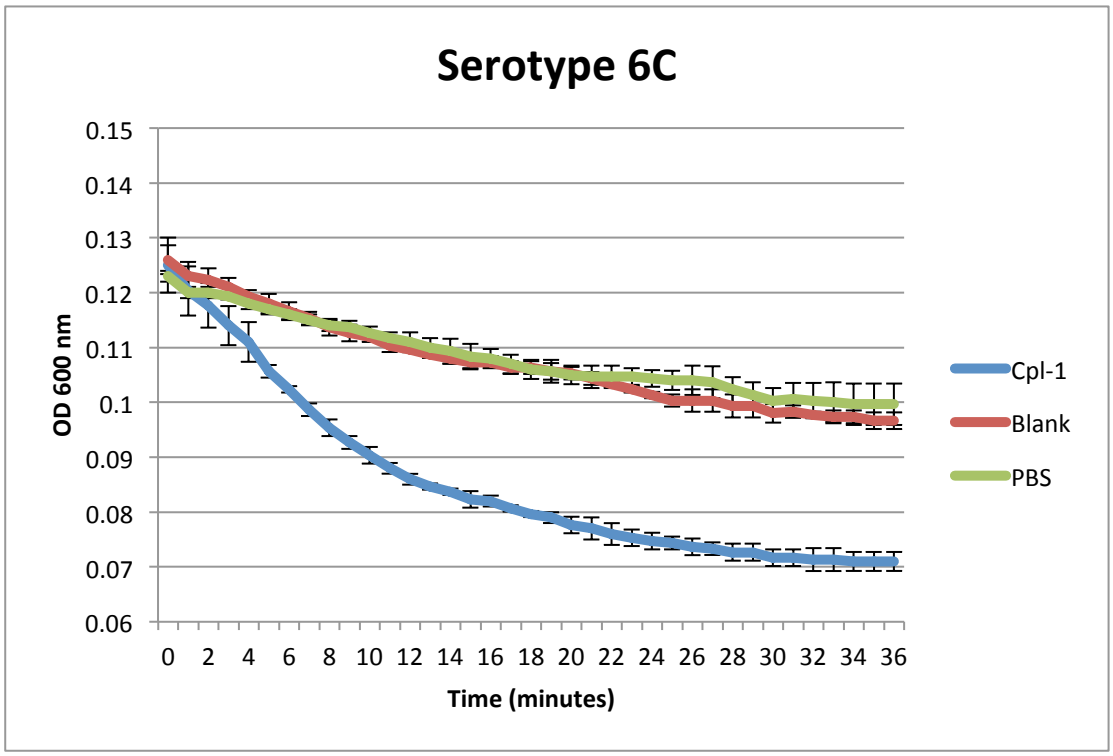
To show clinical relevance for *C. reinhardtii* derived Cpl-1 as a next generation antibiotic, preliminary assays were conducted against a panel of clinical isolates supplied by the Royal Free Hospital, including those with reduced sensitivity to conventional antibiotics (Figure 3.29). Of the four isolates investigated, enriched *C. reinhardtii* Cpl-1 shows definite activity against three serotypes - 6A, 6B, and 6C. Serotype 27 did not appear to show sensitivity to Cpl-1; however, this is by no means conclusive due to the limited scope of the experiment conducted. Further work was again restricted by limited time.

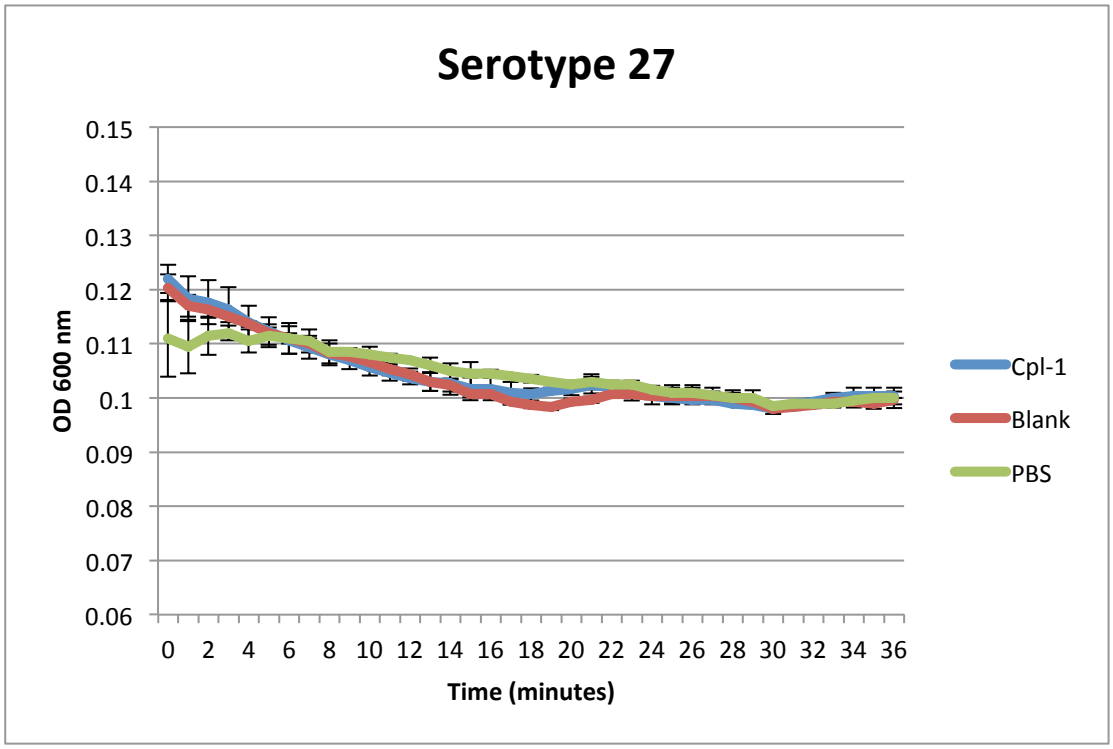
Figure 3.29 – Optical density based clearance activity assay for enriched *C. reinhardtii* derived Cpl-1 against a panel of 4 clinical isolates of *S. pneumoniae* (following two pages)

Serotypes 6A, 6B, and 6C all show activity for Cpl-1 relative to the blank and PBS controls. Serotype 27, however, does not appear to be affected at all by the presence of Cpl-1.


Serotype 6A
OD 600 nm
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (minutes)
Cpl-1
Blank
PBS


Serotype 6B
OD 600 nm
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (minutes)
Cpl-1
Blank
PBS


Serotype 6C
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
OD 600 nm
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (minutes)
Cpl-1
Blank
PBS


Serotype 27
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
0.06
OD 600 nm
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (minutes)
Cpl-1
Blank
PBS

3.2.6.4 ***Specificity of Cpl-1 for* S. pneumoniae**

As discussed above (1.1.4.2), one of the main advantages of endolysins over conventional antibiotics is the high level of specificity conferred by the host phage for the bacterial target. To assess whether *C. reinhardtii* derived Cpl-1 did indeed exhibit such specificities, enriched *C. reinhardtii* Cpl-1 was assessed by clearance assay against two other bacterial species, *E. coli* and *Streptococcus pyogenes*. These data, though not as conclusive as previous results, do not show significant deviation of Cpl-1 containing samples relative to TN72 blank negative controls. This is in agreement with previous reports on Cpl-1 target specificity (Loeffler *et al.*, 2001).

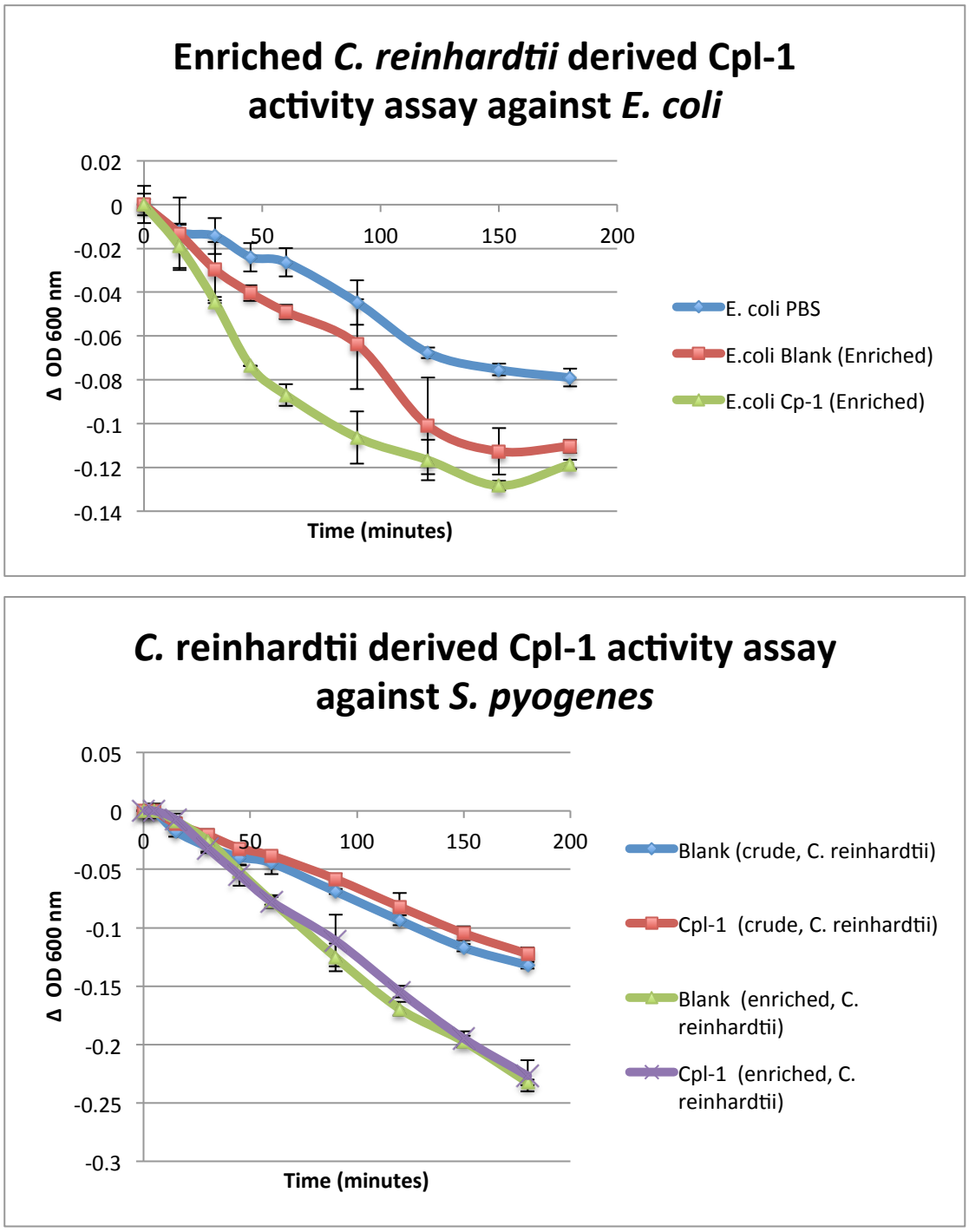


Figure 3.30 – Optical density based clearance activity assay for enriched *C. reinhardtii* derived Cpl-1 against two non-*S. pneumoniae* strains

Neither *E. coli* nor *S. pyogenes,* when treated with *C. reinhardtii* derived Cpl-1, show a significant decrease in optical density relative to a corresponding blank sample.

3.3 Discussion

In the course of this project the potential for expression in the *C. reinhardtii* chloroplast of next generation antibiotics such as Cpl-1 has been realised. The endolysin has been shown to be active and stable, and a robust purification protocol has been developed (Figure 3.31). In the course of achieving these objectives several other avenues have been explored, both in relation to Cpl-1, and to the use of the *C. reinhardtii* chloroplast platform in more general terms.

3.3.1 Comparison of the pASap1 and pSRSap1 expression systems

As discussed in the introduction to this thesis, there is a growing interest in the use of *C. reinhardtii* as a platform for recombinant protein expression. However, the achievement of such goals is hindered by low levels of expression. During the course of this chapter the novel expression vector pSRSap1 that uses the *psaA* promoter and 5' UTR has been investigated alongside the older pASap1 vector, which drives transgenes off the *atpA* expression signals. Direct comparison of Cpl-1 yields when driven off *psaA*, relative to *atpA* show between a 1.5 and 2 fold increase in Cpl-1 accumulation. This increase in productivity is not enough to allow *C. reinhardtii* to compete with the mature recombinant expression platforms (Finnis *et al.*, 2010; Li *et al.*, 2011; Oey *et al.*, 2009a; Swiech *et al.*, 2012), however it does represent an incremental step towards this objective.

The next step in this process can be seen as understanding how this increase in product accumulation was achieved, so to then improve upon it. By comparing expression of the same gene under the two different promoters/ 5' UTRs we can be confidant that this increase in yield is unrelated to protein turnover; however, what is less clear is whether the limiting factor is at the level of transcription or translation. Answering this question is important for directing how to further optimise future *C. reinhardtii* chloroplast expression cassettes and could be addressed fairly simply by transcript analysis using northern blot or RT-PCR. Due to time constraints this work is yet to be carried out, but given previous work on chloroplast gene expression regulation it is likely that control is being exerted at a translational level, and as such transcript accumulation would likely be in excess in either system (Coragliotti *et al.*, 2011).

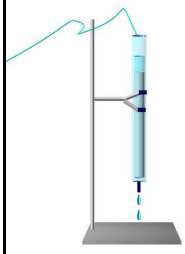
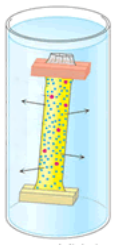
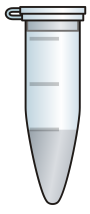
	A *C. reinhardtii* strain expressing *cpl-1* is grown to mid-late log phase and equalized to ~100x culture concentration
	Cells are broken by pressure cell disruption. Extract is incubated on ice for 30 minutes to allow aggregation of membrane fragments.
	Crude cell extract is diluted to ~10x culture concentration and centrifuged.
	Soluble protein extract is enriched by ion exchange chromatography. Column is washed with 1.5 M NaCl and Cpl-1 specifically eluted with choline chloride.
	Resulting elutant is dialyzed overnight to remove bound choline.
	Enriched sample is split into aliquots and stored at -80 °C

Figure 3.31 – Production of enriched Cpl-1 preparation

Each stage of the Cpl-1 enrichment process is illustrated above with a simplified description of techniques used. For further details see section 2.5.

Further investigation into improving expression from this vector would hence be focused on 5' UTR optimisation, in particular looking into and removing potential auto-attenuation based regulators, as has been reported recently by others (Specht and Mayfield, 2013).

It is likely that such investigations would ultimately lead back to the nucleus (Stern and Harris, 2009). The beauty of the chloroplast is that it can be viewed as a pseudo-bacterium with eukaryotic benefits. It must not be forgotten however that these benefits come at the cost of tight genetic control from the nucleus which must be overcome, or at least mediated, before the potential for *C. reinhardtii* chloroplast based recombinant protein production can be fully realised (Manuell *et al.*, 2007; Specht and Mayfield, 2013).

3.3.2 Issues of Cpl-1 solubility in *C. reinhardtii* chloroplast protein preparations

The issues encountered relating to maintaining Cpl-1 in solution when producing concentrated protein preparations were unexpected. Previous reports from other groups where Cpl-1 was produced in *E. coli* had not discussed such an issue, nor had other groups who had synthesised recombinant proteins in the *C. reinhardtii* chloroplast. At the time when the issue was first encountered recombinant protein expression in the Purton lab had been limited to chemically broken and solubilised extracts, meaning there were no past experiences to attest to the viability of mechanical cell breakage. As the lab moved more towards the expression of recombinant proteins it became apparent that the solubility problems observed for Cpl-1 were indeed unique. Research conducted by others has showed the retention of several recombinant proteins in solution even at high cell concentrations.

Further investigation has shown Cpl-1 synthesised in *E. coli* to be soluble, and the similar lysin Pal (which also binds choline) to be soluble when synthesised in the *C. reinhardtii* chloroplast. It has also been shown that once *C. reinhardtii* synthesised Cpl-1 is purified it can then be concentrated without further precipitation. The issue is hence unique to *C. reinhardtii* synthesised Cpl-1 and only occurs when in high concentration crude cell extracts, however it does not seem to be connected to the choline binding element of the enzyme.

It thus seems likely that the issue is caused by interaction of Cpl-1 with some (most likely insoluble) element present in the *C. reinhardtii* cell extract. We know this interaction to be concentration related and also reversible due to experiments where cell breakage has been conducted at high concentration prior to dilution of extracts for centrifugation. It is also likely to involve a component of particulate as opposed to membrane nature, as evidenced by the lost of Cpl-1 to the pellet at low centrifugation speeds incapable of removing membranes from solution. In addition surface hydrophobicity analysis of Cpl-1 has failed to reveal any areas likely to be membrane interacting. The possibility of interaction with the choline binding sites was considered (despite the absence of such issues for Pal), but incubation with excess free choline chloride (blocking the binding site) failed to increase Cpl-1 solubility. The protein-based cell wall of *C. reinhardtii* was also considered as a possible interaction partner; however, the persistence of the problem in the cell wall deficient background of TN72 suggested this not to be the case.

Ultimately it is still unclear why Cpl-1 precipitates at high cell concentrations, but this issue has been mainly resolved by the centrifugation of broken cells at a cell concentration not exceeding 10x culture volume, or 2 x 10^8 cells/ml. A possible route for further investigation could be the addition of enriched *C. reinhardtii* or *E. coli* synthesised Cpl-1 to concentrated wild type *C. reinhardtii* cell extracts to see if this could induce precipitation. If loss of Cpl-1 solubility was observed then this would be further evidence of direct interaction with a *C. reinhardtii* cellular component, and further research could be conducted to elucidate the nature of such a moiety.

3.3.3 Considerations relating to non-denaturing cell breakage techniques

Prior to the initiation of this project, little work had been conducted in the Purton lab on therapeutic protein expression in *C. reinhardtii*. As such, there was limited experience to draw on for the production of non-denatured *C. reinhardtii* protein preparations. Mechanical cell breakage of the cell walled transformant line was only possible by pressure cell disruption or sonication; however, the creation of the cell wall deficient lines allowed for a range of further breakage techniques to be investigated. On examining the relative efficiencies of the new methods available (3.2.4.2), it was shown that a pressure-based disruption method was still

the most effective in terms of product harvest, and thus was selected for further use. It should be noted however, that the other breakage techniques, though less efficient, do have merits that could make them superior to pressure cell disruption in certain situations.

On a lab-scale the use of a cell disrupter is feasible as the volumes to be disrupted are generally in the 5 - 50 ml range. If production of therapeutic proteins were to be conducted on an industrial scale, with disruption volumes measured in litres as opposed to millilitres, an alternative method of cell breakage would be required. Considering the data presented in Figure 3.17 and Chart 3.6, one technique of particular interest is the bursting of cells by osmotic shock using ddH_2O. In the experiment conducted, such a treatment yielded only 10 % of the Cpl-1 seen for the cell disrupted sample; however, the dramatic difference in labour and equipment costs between the two techniques must be considered, assuming such a low osmotic potential is not deleterious to the protein itself. Taking this into account there is certainly an argument for wider investigations into the effects of higher dilutions over prolonged periods of time.

At lab-based scale, treatment with mild detergent would be another cell breakage method worth further investigation. Assuming cell breakage can be achieved below the critical micelle concentration of the detergent, interactions with an ion exchange column should be minimal and thus the addition of detergent could be a very rapid method of protein extract preparation for purification. Further investigation into minimum levels required for efficient cell breakage should thus be considered.

Finally, on grounds of simplicity and sterility it is worth optimising the freeze-thaw breakage technique. This is of particular interest as current practice in the Purton lab is for therapeutic protein expressing lines to be grown on a large lab scale of 30 L, and harvested biomass stored at -80 °C until needed. A system where samples could be thawed in a pre-lysed state would thus be an advantage.

As the Purton lab seems to be progressively turning towards the expression of protein therapeutics in the *C. reinhardtii* chloroplast, and TN72 is now the primary

recipient line, it is likely that each of these proposals will in due course be investigated fully.

3.3.4 Observations on Cpl-1 purification

Isolation of Cpl-1 was significantly simplified by the availability of a proven purification regime (Loeffler *et al.*, 2001). This method was shown (with some modifications to the protocol) to be highly effective, and thus was chosen as the primary purification method employed. In regard to moving the project forward the main issue identified while working with this protocol has been the contaminating proteins found in the elution samples. For the purposes of this project these were not considered to be an issue so further action was not taken, however, if the project progresses to an *in vivo* platform then a true degree of purification (as opposed to enrichment) would be required.

The most apparent means of increasing the level of purification would be to increase the ionic potential and possibly volume of the wash stages. As shown previously, no detectable Cpl-1 is released from the column during the current wash stages, suggesting that the NaCl concentration could be increased without prematurely eluting the protein of interest. The first step in optimising the process would be to elute Cpl-1 with a salt gradient to see how high a salt concentration could be tolerated before Cpl-1 started to elute. Further optimisation could be conducted from there.

Alternatively the ion exchange protocol could be left as is, and a polishing step added. This could be achieved with an HA affinity column to give a highly pure final product. The disadvantage of such a technique however is the high cost of such a column, especially once the process is scaled up. Nickel affinity chromatography however is considerably cheaper and could be used as a single stage purification with the replacement of the current HA tag with a polyhistidine tag, although further activity assays would be required to ensure activity was not compromised. Gel exclusion chromatography was briefly investigated as a polishing technique and although promising results were not obtained, this does not rule out future use.

3.3.5 The potential of *C. reinhardtii* synthesised Cpl-1 as a next generation antibiotic

The primary aim of this project was to demonstrate that active Cpl-1 could be produced in the *C. reinhardtii* chloroplast, and by extension open the doors for further lysins to be investigated in this platform. This has been achieved: Cpl-1 accumulation has been demonstrated, it has been shown to be specifically active against the target pathogen, and a purification strategy demonstrated. It must now be considered whether this project should be seen as a proof-of-concept alone, or if there is potential for Cpl-1 to be investigated further, and whether there is ultimately space in the market for such a product. In either case, various experiments are required to properly conclude the work; the most pressing being the quantification of activity by more sophisticated enzymatic analysis, and the further characterisation of Cpl-1 target range by screening activity against a broader range of *S. pneumoniae* strains and other related bacteria.

As discussed at length in the Chapter one (1.1.1), the pivotal factor in the dwindling supply of effective antibiotics is not the rise of antibiotic resistance – as this has always been an on-going process – but the decline of novel therapeutic agents. This can partly be seen as an issue of depletion, the 'low hanging fruit' having already been taken; however, the consolidation of pharmaceutical companies and an industry-wide shift away from antibiotic development has also played a major part (Payne *et al.*, 2007). This is understandable. Development of therapeutic products is expensive, and a product that carries a substantial risk of becoming obsolete due to resistance development can be seen as a hazardous investment. Taking this into account, the question over the potential of Cpl-1 becomes not 'would it work therapeutically?' but 'would it work economically?'

The lysins show a remarkable opposition to resistance development, which helps to lower the risk associated with an antibiotic product. However, one of the main strengths of Cpl-1 as a therapeutic, namely its high level of specificity, can also be viewed as a distinct economic disadvantage. A highly specific product has to go though the same regulatory stages as a broad-spectrum agent, but on reaching the clinic has a far smaller market from which to recoup the expense. Additionally, specific diagnosis of not only the presence of the target organism, but also

attribution with disease is required, reducing the target market even further. At present these barriers to market are likely to be too high for the investment required to get *C. reinhardtii* derived Cpl-1 into the clinic, even with the benefits of a low cost algal expression platform.

With time, however, this is liable to change for two reasons. Firstly the therapeutic landscape is likely to shift. As effective last-resort antibiotics become scarcer, the demand for next generation antibiotics will increase. This is likely to be aided by the advancement of research into the human micro-biome and pathologies relating to disruption of the natural bacterial flora. Developments are also likely in the field of lysin biotechnology. The majority of lysins currently under investigation (including Cpl-1) can be classed as first generation, or native lysins. The next step in lysin development will be the application of synthetic biology to the native lysins to create enzymes with, for example, broader pathogenic host ranges, improved catalytic activity, and improved bioavailability and immune tolerance (Schmelcher *et al.*, 2012).

Longer term further work on *C. reinhardtii* synthesised Cpl-1 will thus fall into two categories: investigations into improving the therapeutic properties of the enzyme, and continuing the use of native Cpl-1 as a proof of concept by moving into *in vivo* efficacy trials. Both of these avenues are discussed further in Chapter six.

Chapter 4

Attempts to express other lysins and a study into the problems of foreign gene expression in the *C. reinhardtii* chloroplast

4.1 Introduction

4.1.1 Production of further lysins in the *C. reinhardtii* chloroplast

After the successful expression of the *S. pneumoniae* lysin Cpl-1 (see Chapter three), two further lysins were selected for expression in the *C. reinhardtii* chloroplast. The general suitability for lysin expression in the *C. reinhardtii* chloroplast has previously been discussed; however, to emphasize the potential of *C. reinhardtii* as an expression platform, care was taken to select targets able to exploit the unique advantages associated with this organism.

4.1.1.1 *Lys16 from the* Staphylococcus aureus *phage P68*

4.1.1.1.1 Background to *S. aureus* as a therapeutic target

Staphylococcus aureus is a coccus Gram-positive facultative anaerobe that is associated with a large variety of human interactions. These range from commensal colonization of the skin and nasal pharynx, to wound infection, to fatal bacteremia and meningitis (Kluytmans *et al.*, 1997). As with *S. pneumoniae*, *S. aureus* is considered to be a re-emerging pathogen, illustrated by the huge media awareness of methicillin resistant *S. aureus* (MRSA) as the quintessential 'superbug'. It is estimated that in 2005, MRSA killed over 18,000 people in the USA: more than the number of deaths attributed to AIDS. As well as methicillin resistance, strains of *S. aureus* have now been isolated with resistance to vancomycin and quinolone based therapies, severely limiting treatment options (Lowy, 2003). Antibiotic resistant strains are also increasingly being seen in the community where previously they were confined to healthcare facilities, an example being the highly virulent *USA300* (Leclercq, 2009). There is clearly a pressing need for novel antibiotics to fight *S. aureus*; however, this organism's long track record of resistance development indicates the need for a treatment that can oppose this process. This is clearly an area where lysins could prove to be highly suitable.

An additional factor that makes lysin therapy particularly applicable to *S. aureus* is activity on mucosal membranes. The most common transmission pathway of *S. aureus* in hospitals is naso-orally from the nasal-pharynx reservoir of carriers (von

Eiff *et al.*, 2001), and thus mucosal targeting antibiotics such as mupirocin are routinely used to clear this reservoir in new patients (Coates *et al.*, 2009). This practice has been shown to be highly effective in slowing the spread of infection, but resistant strains are rapidly emerging, and the rarity of mucosal membrane targeting antibiotics means there are few in place to replace mupirocin when resistance becomes widespread.

4.1.1.1.2 The history of *S. aureus* lysins

The first reported *S. aureus* lysin was identified by Ralston and McIvor in 1955 from the phage K1 (Ralston *et al.*, 1955). This lysin (termed virolysin), however, was apparently unable to lyse cells unless they were 'sensitised' first, by either the presence of the phage itself or by physical treatments such as acetone, UV irradiation, or heating to 56 °C. The first 'stand alone' lysin specific to *S. aureus* was identified in 1971 by Sonstein *et al* (Sonstein *et al.*, 1971). The Phage Associated Lysin (PAL, not to be confused with the *S. pneumoniae* lysin Pal) isolated was shown to lyse all staphylococcal stains tested without affecting other bacterial species, as well as being capable of degrading purified staphylococcal cell walls. It was noted that the enzyme could be used to create spheroplasts when applied in the presence of 7.5 % PEG 4000 (to provide osmotic stability), and relevant applications to the molecular biology of *S. aureus* were suggested. As with early work on Cpl-1 however, no mention was made as to the potential of PAL as an antimicrobial. Since then, *S. aureus* lysins have become one of the most populated areas of the field, with numerous lysins being characterised and several structural rearrangements investigated including truncations (Horgan *et al.*, 2009) and chimeric enzymes (Manoharadas *et al.*, 2009).

4.1.1.1.3 The *S. aureus* phage P68, and the Lys16 lysin

The lytic properties of the *S. aureus* phage P68 lysin Lys16 was first investigated by Manoharadas and colleagues (Manoharadas *et al.*, 2009). Local sequence alignment analysis using the Blastp platform showed sequence identity to the related *S. aureus* lysins Twort and Φ11; however, whereas both such enzymes show a three-domain structure featuring two distinct catalytic domains, Lys16 contains a single N-terminal D-alanyl-glycyl endopeptidase domain and the archetypal C-terminal

cell wall binding region. Antimicrobial activity was confirmed although the work was not followed up due to reported issues surrounding protein solubility in the *E. coli* expression system. Instead the group proceeded to produce a chimera of the Lys16 N-terminal catalytic domain with the C-terminal cell binding domain of p17, a viron bound lysin from the same phage thought to be involved in initial infection of the bacterial host. This produced a soluble protein and was demonstrated to have lytic activity against *S. aureus*.

4.1.1.1.4 Suitability of *C. reinhardtii* as an expression platform

As a proof of concept regarding the enhanced stability and protein folding ability of the *C. reinhardtii* chloroplast relative to its prokaryotic cousins, the unmodified *lys16* gene was expressed with the hope of showing superior levels of solubility relative to the *E. coli* system.

4.1.1.2 ***The putative lysin* gp20 *from the* Propionibacterium acnes *phage PA6***

4.1.1.2.1 Background to *P. acnes*

Propionibacterium acnes is a common human skin commensal which is thought to be the main causative agent of inflammatory acne vulgaris as well as several other opportunistic infections of the upper dermal layers. It is an aerotolerant anaerobic Gram-positive bacillus, which is often found as part of the natural skin flora predominately on lipid rich areas such as the face, neck, and shoulders (Farrar *et al.*, 2007). In certain situations *P. acnes* can shed its commensal persona and pathogenically infect pores leading to the inflamed lesions characteristic of acne vulgaris. At present the precise mechanism for the activation of the inflammatory response is unclear, although two dominant theories persist. There is evidence that indicates specific adaptive activation of CD4+ T-cells via antigenic processing from Langerhans cells in the follicle cell wall, triggering inflammation. It has also been shown that *P. acnes* stimulates the immune system via the innate pathway by inducing production of pro-inflammatory cytokines such as Il-1α and Tumour Necrosis Factor (TNF) in keratinocytes. These recruit monocytes to the follicle, which themselves release pro-inflammatory cytokines. The initiating factor in either case has been shown to be the increase in *P. acnes* population density at the follicle (Farrar and Ingham, 2004). Once the inflamed lesion is established it can

provide an entry point for other opportunistic pathogens such as the aforementioned *S. aureus.*

This condition affects approximately 80 % of individuals (most commonly during adolescence), with lesions resulting in permanent scarring in around 30 % of cases (Farrar *et al.*, 2007). Acne vulgaris typically clears spontaneously by the late teens or early twenties, although more severe cases require therapeutic intervention. This is also the case for patients of an immunocompromised nature, where the presentation of novel pathogenic entry sites is of particular concern. Standard treatment falls into two categories: the use of broad-spectrum antibiotics such as cephalosporin or erythromycin, administered either orally or topically (Brook and Frazier, 1991), and the topical application of harsh chemical agents such as benzyl peroxide. Neither treatment is ideal; chemical treatment carries various side effects, and the nature of the infection lends itself to antibiotic resistance development, as both topical and systemic administration results in a concentration gradient forming across the upper layers of the dermis. Coupled with the highly mobile nature of many relevant antibiotic resistant genes, this results in rapid generation of resistant strains. An example of such a mobile element is the 23S ribosomal methylase, which, via target site modification, can simultaneously give resistance to macrolide, lincosamide and type B streptogramin classes of antibiotics (Eady *et al.*, 1989).

4.1.1.2.2 Applicability of lysin based treatment, and identification of a suitable lysin

Issues of rapid resistance development and exposure of the patient to opportunistic fungal infections due to clearance of natural skin flora indicate a definite and as of yet unexploited opportunity for a lysin based therapy. To date fourteen *P. acnes* phages have had their full genomes sequenced; the first in 2007 (Farrar *et al.*, 2007), with two more in 2011 (Lood and Collin, 2011), and a further eleven in 2012 (Marinelli *et al.*, 2012). As noted by Marinelli *et al* there is a remarkable lack of diversity between *P. acnes* phage isolates, even between those where isolation was separated by an excess of 30 years. Each of these fourteen phage genomes contains a putative lysin (identified as such by Blastp and position within the viral genome) encoded by ORF20; however, to date endolytic activity has yet to be confirmed.

4.1.1.2.3 Potential for a cosmetic route to market

Application of a lysin-based therapy targeting *P. acnes* could have enormous value in treating clinical cases of acne vulgaris, but also holds great potential for the cosmetic industry in the treatment of non-clinical cases of acne. Classification as a cosmetic rather than a pharmaceutical product significantly reduces regulatory hurdles, and simultaneously opens a significantly larger market due to the widespread nature of mild to moderate cases of acne, particularly amongst adolescents. The skin-care product market is already well established and collaboration with an existing player could potentially allow for drop-in entry to the industry. Such factors would enormously ease the transition of such a product from the lab to the marketplace.

4.1.1.2.4 Suitability of *C. reinhardtii* as an expression platform

Due to issues of bioavailability and immune clearance associated with protein-based therapies, a topical application would clearly be most appropriate. *C. reinhardtii* is thus perfectly suited as an expression platform as the organism's GRAS status would require only precursory purification of the lysin, although a DNase step would be required to remove all genetically modified DNA. In line with current trends in the industry the resulting product could easily be marketed as a botanical extract. This benefit is in addition to factors of host adaptation and lack of native substrate discussed previously. For this study the putative lysin gene *gp20* from phage PA6 (the first, and at the time only, *P. acnes* phage genome to be sequenced) was chosen for expression in the *C. reinhardtii* chloroplast.

4.1.2 A note on figure presentation

Within this chapter numerous different genes are expressed from a range of vectors in several different organisms. To mitigate any confusion and allow for rapid identification of figures, a graphical system taking inspiration from ancient Egypt has been derived. This system is fully explained in Figure 4.1.

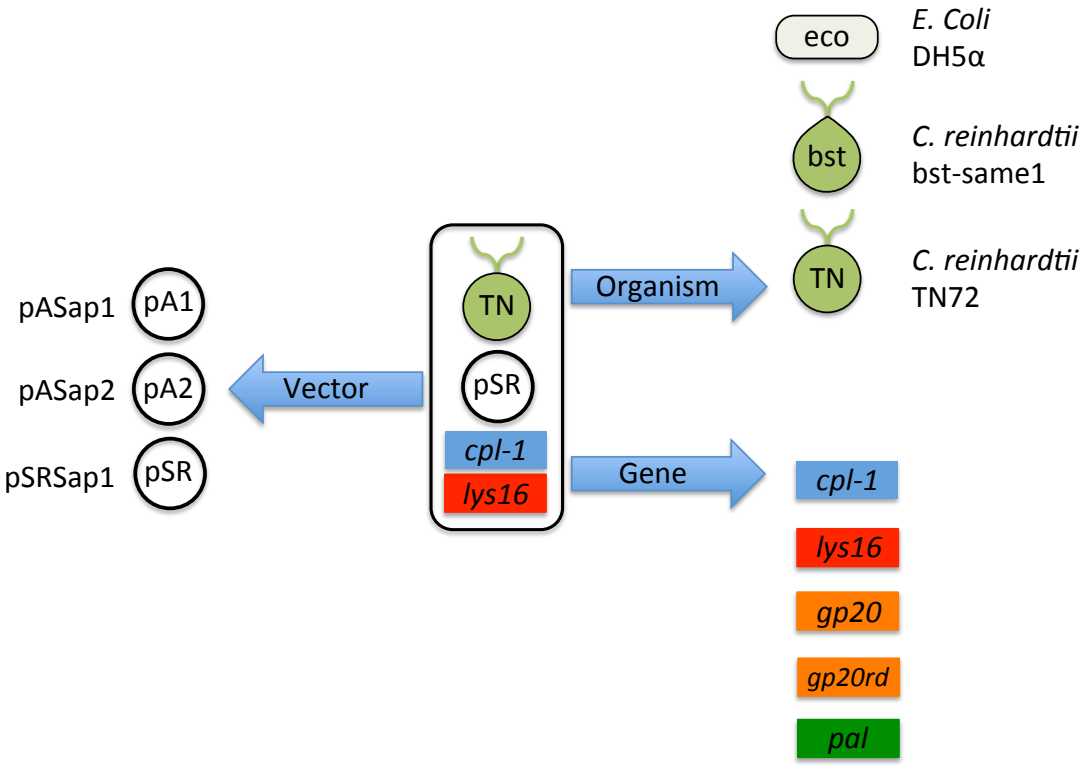


Figure 4.1 – An illustrated key to graphical figure labels in this chapter

Each cartouche represents the construct being analysed in the figure it accompanies. It is formed of three parts, the upper, middle, and lower sections representing the organism and strain, the vector, and the gene of interest, respectively. In the case of fusion genes, both gene symbols are displayed, with the upstream gene above the downstream as shown in this example for pSRSap.cpl-1:lys16 in *C. reinhardtii* TN72.

4.1.3 Aims and objectives

The aims of the chapter were as follows

- To express the *Staphylococcus aureus* phage P68 lysin *lys16* in the *C. reinhardtii* chloroplast, and investigate any improvement of solubility seen over *E. coli* derived product.

- To express the *P. acnes* phage PA6 putative lysin *gp20* in the *C. reinhardtii* chloroplast, confirm its activity, and investigate the suitability of *C. reinhardtii* derived *gp20* as a cosmetic product.

As is presented in this chapter, the expression of both genes in the *C. reinhardtii* chloroplast proved to be problematic, and as such the following aims were added:

- To investigate factors affecting the translation initiation of non-detectable recombinant proteins in the *C. reinhardtii* chloroplast.

- To investigate the effect of the novel gene design software, the Codon Usage Optimizer, on the expression of recombinant proteins in the *C. reinhardtii* chloroplast.

4.2 Results

4.2.1 Initial transformation with *lys16* and *gp20* utilizing the pASap1 vector

4.2.1.1 *Selection and design of genes*

4.2.1.1.1 Confirmation of *gp20* as an endolysin

Farrar *et al* had previously postulated that the putative ORF *gp20* identified in the PA6 genome encodes the endolysin (Farrar *et al.*, 2007), but steps were taken here to verify their conclusions. Investigations into the positioning of *gp20* on the genome show it to be located directly upstream of a putative class II holin gene; this is typical of lysin gene arrangement (Figure 4.2). A local sequence alignment analysis using the Blastp platform was conducted next. A conserved domain analysis (Figure 4.3a) identified an *N*-terminal amidase domain, and revealed common patterns related to the peptidoglycan recognition protein (PGRP) superfamily. A comparative Blastp (Figure 4.3b) revealed the top 15 hits to be from *gp20*s of other *P. acnes* phage, illustrating the high degree of sequence identity previously observed (Marinelli *et al.*, 2012) between these viruses. The next hit relating to an annotated protein is a *P. acnes* N-acetylmuramolyl-L-alanine amidase, a bacterial enzyme involved in peptidoglycan rearrangement. Clustal Omega alignment analysis with this protein revealed a high degree of similarity for the Gp20 N-terminal catalytic domain, juxtaposed by a complete absence of sequence identity for the C-terminal region (Figure 4.4). This is typical of Gram-positive lysin structure.

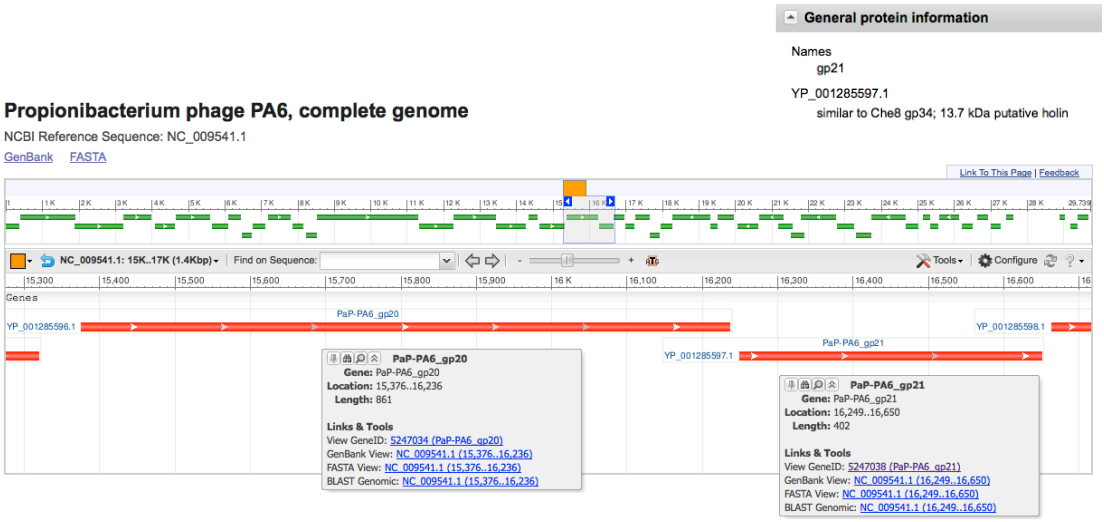
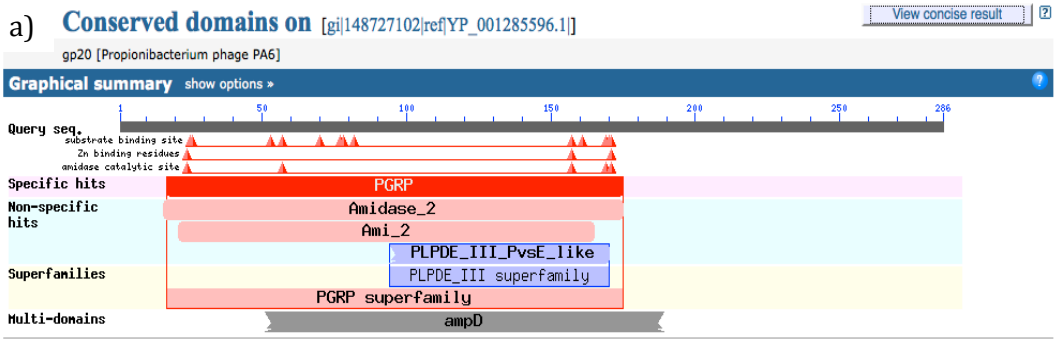


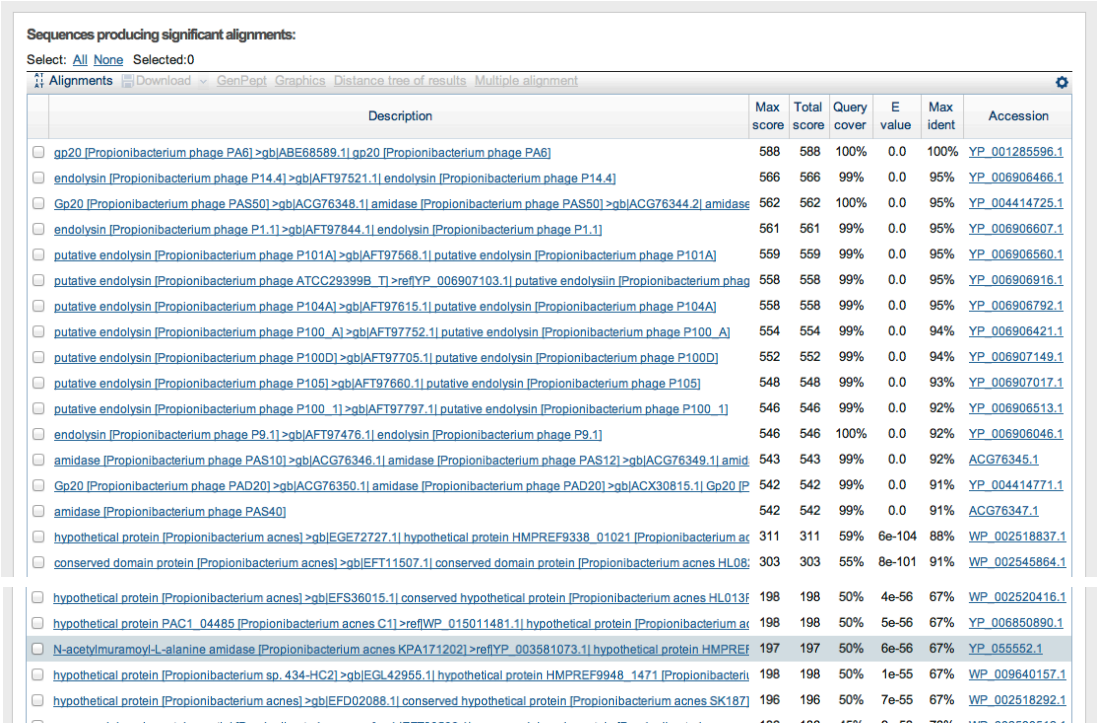
Figure 4.2 – Graphical illustration of the PA6 genome showing gene mapping of *gp20* and the putative holin *gp21*

The putative lysin *gp20* is shown to be located directly upstream of *gp21*, a putative holin. This positioning is typical of lysin:holin gene arrangement.

a)



b)



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
gp20 [Propionibacterium phage PA6] >gb|ABE68589.1| gp20 [Propionibacterium phage PA6]	588	588	100%	0.0	100%	YP_001285596.1
endolysin [Propionibacterium phage P14.4] >gb|AFT97521.1| endolysin [Propionibacterium phage P14.4]	566	566	99%	0.0	95%	YP_006906466.1
Gp20 [Propionibacterium phage PAS50] >gb|ACG76348.1| amidase [Propionibacterium phage PAS50] >gb|ACG76344.2| amidase	562	562	100%	0.0	95%	YP_004414725.1
endolysin [Propionibacterium phage P1.1] >gb|AFT97844.1| endolysin [Propionibacterium phage P1.1]	561	561	99%	0.0	95%	YP_006906607.1
putative endolysin [Propionibacterium phage P101A] >gb|AFT97568.1| putative endolysin [Propionibacterium phage P101A]	559	559	99%	0.0	95%	YP_006906560.1
putative endolysin [Propionibacterium phage ATCC29399B_T] >ref|YP_006907103.1| putative endolysiin [Propionibacterium phag	558	558	99%	0.0	95%	YP_006906916.1
putative endolysin [Propionibacterium phage P104A] >gb|AFT97615.1| putative endolysin [Propionibacterium phage P104A]	558	558	99%	0.0	95%	YP_006906792.1
putative endolysin [Propionibacterium phage P100_A] >gb|AFT97752.1| putative endolysin [Propionibacterium phage P100_A]	554	554	99%	0.0	94%	YP_006906421.1
putative endolysin [Propionibacterium phage P100D] >gb|AFT97705.1| putative endolysin [Propionibacterium phage P100D]	552	552	99%	0.0	94%	YP_006907149.1
putative endolysin [Propionibacterium phage P105] >gb|AFT97660.1| putative endolysin [Propionibacterium phage P105]	548	548	99%	0.0	93%	YP_006907017.1
putative endolysin [Propionibacterium phage P100_1] >gb|AFT97797.1| putative endolysin [Propionibacterium phage P100_1]	546	546	99%	0.0	92%	YP_006906513.1
endolysin [Propionibacterium phage P9.1] >gb|AFT97476.1| endolysin [Propionibacterium phage P9.1]	546	546	100%	0.0	92%	YP_006906046.1
amidase [Propionibacterium phage PAS10] >gb|ACG76346.1| amidase [Propionibacterium phage PAS12] >gb|ACG76349.1| amid	543	543	99%	0.0	92%	ACG76345.1
Gp20 [Propionibacterium phage PAD20] >gb|ACG76350.1| amidase [Propionibacterium phage PAD20] >gb|ACX30815.1| Gp20 [P	542	542	99%	0.0	91%	YP_004414771.1
amidase [Propionibacterium phage PAS40]	542	542	99%	0.0	91%	ACG76347.1
hypothetical protein [Propionibacterium acnes] >gb|EGE72727.1| hypothetical protein HMPREF9338_01021 [Propionibacterium ac	311	311	59%	6e-104	88%	WP_002518837.1
conserved domain protein [Propionibacterium acnes] >gb|EFT11507.1| conserved domain protein [Propionibacterium acnes HL08	303	303	55%	8e-101	91%	WP_002545864.1
hypothetical protein [Propionibacterium acnes] >gb|EFS36015.1| conserved hypothetical protein [Propionibacterium acnes HL013	198	198	50%	4e-56	67%	WP_002520416.1
hypothetical protein PAC1_04485 [Propionibacterium acnes C1] >ref|WP_015011481.1| hypothetical protein [Propionibacterium ac	198	198	50%	5e-56	67%	YP_006850890.1
N-acetylmuramoyl-L-alanine amidase [Propionibacterium acnes KPA171202] >ref|YP_003581073.1| hypothetical protein HMPREF	197	197	50%	6e-56	67%	YP_055552.1
hypothetical protein [Propionibacterium sp. 434-HC2] >gb|EGL42955.1| hypothetical protein HMPREF9948_1471 [Propionibacteri	198	198	50%	1e-55	67%	WP_009640157.1
hypothetical protein [Propionibacterium acnes] >gb|EFD02088.1| conserved hypothetical protein [Propionibacterium acnes SK187]	196	196	50%	7e-55	67%	WP_002518292.1

Figure 4.3 – Conserved domain analysis (a) and comparative Blastp search (b) of the *P. acnes* phage PA6 putative lysin Gp20

A conserved domain analysis **(a)** shows an N-terminal amidase region as well as peptidoglycan recognition protein (PGRP) motifs: both hallmarks of a lysin.

Comparative sequence alignment using the Blastp platform **(b)** shows the highest levels of sequence identity to be from putative lysins of other *P. acnes* phages. After these, the next annotated protein is an N-acetylmuramoyl-L-alanine amidase, again supporting the case that Gp20 is a lysin.

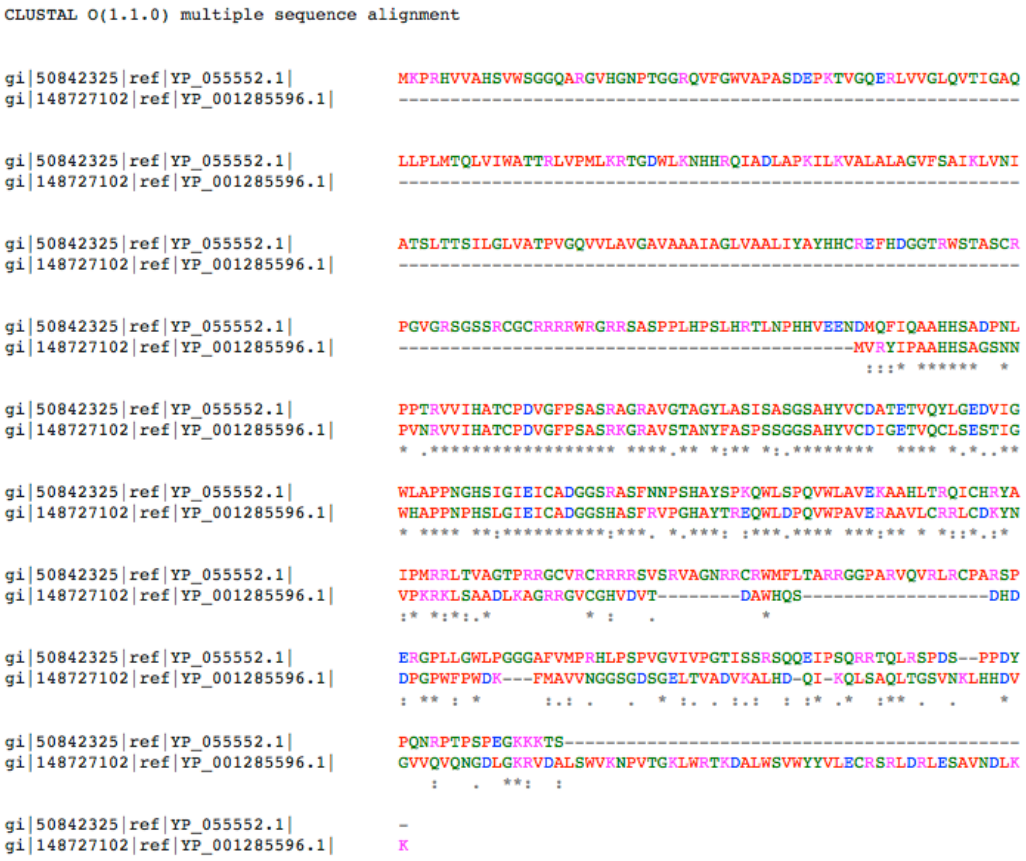


Figure 4.4 – Clustal Omega alignment analysis of the *P. acnes* phage PA6 putative lysin Gp20 (lower line) with the N-acetylmuramoyl-L-alanine amidase identified in Figure 4.4 (upper line)

The PA6 Gp20 N-terminal domain shows high homology with the N-acetylmuramoyl-L-alanine amidase, which is then lost at the C-terminus. This supports the standard lysin domain structure of N-terminal catalytic and C-terminal cell wall binding domains.

4.2.1.1.2 Gene design and synthesis

Protein sequences for PA6 Gp20 and P68 Lys16 were acquired from Genbank, accession numbers YP_001285596 and NP_817310, respectively. Both genes were synthesised by GeneArt[12] with the addition of 5' *Sap*I and 3' *Sph*I sites, and gene sequence encoding a C-terminal Haemagglutinin (HA) epitope tag. The genes were codon optimised to a Codon Adaptation Index (CAI) of 0.8 using the Kazusa CAI table[13] (Nakamura *et al.*, 2000), with the explicit avoidance of *Sap*I and *Sph*I sites.

4.2.1.2 ***Creation of* lys16 *and* gp20 *containing* C. reinhardtii *lines***

Both the *gp20* and *lys16* genes were cloned into the pASap1 vector as described for *cpl-1* above (3.2.1.2), such that expression of each gene would be driven by the *atpA* promoter/5' UTR. Confirmation of insertion into the vector was by colony PCR (**Appendix f**, **Appendix g**) and the coding sequence verified by DNA sequencing using internal overlapping primers (see **Appendix t** for details).

The constructs pASap1.gp20 and pASap1.lys16 were used to transform the *C. reinhardtii* recipient line TN72 via the glass bead method as described (2.3.2.2). As with pASap1.cpl-1 transformation by glass beads, transformant numbers were low, yielding 8 and 14 colonies from pASap1.gp20 and pASap1.lys16, respectively. Putative transformants were screened by PCR (Appendix h, Appendix i) and confirmed by DNA sequencing. Two confirmed lines for each lysin were named LysTGi and LysTGii for *gp20* and LysTLi and LystLii for *lys16,* following the naming system defined in Chapter three (Figure 3.2).

4.2.1.3 ***Expression of* gp20 *and* lys16**

4.2.1.3.1 Expression in *E. coli* strain DH5α

Expression of transgenes in *E. coli* was analysed both to confirm correct construct assembly, and also to provide a comparison between expression in the bacterial and algal platforms. Lines of *E. coli* carrying the plasmids pASap1.cpl-1, pASap1.gp20, pASap1.lys16 and pASap1.empty were prepared for analysis as described (2.4.3). Western blot analysis was conducted followed by immuno-

[12] http://www.invitrogen.com/geneart

[13] http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3055.chloroplast&aa=1&style=N

detection using anti-HA antibodies and visualisation using the LiCor Odyssey system. Figure 4.5 shows low accumulation of Lys16 relative to the Cpl-1 positive control, but no detectable HA-tagged protein at all for Gp20. The low expression of *lys16* and absence of detectable expression for *gp20* was thought to be due to the non-optimised properties of the *atpA* promoter and 5' UTR in *E. coli.*

4.2.1.3.2 Expression in the *C. reinhardtii* chloroplast

Transformed *C. reinhardtii* lines containing *gp20* (LysTGi and LysTGii) and *lys16* (LysTLi, and LysTLii) were prepared for protein analysis as described (2.4.1). Samples were analysed for recombinant protein as for *E. coli* above; however, Figure 4.6 shows no detectable expression for either *lys16* or *gp20*. The strong band seen for the positive control, *cpl-1*, and the complete absence of correct bands for either of the *lys16* or *gp20* samples implies expression levels of a least two orders of magnitude below that of *cpl-1.*

The sequence analysis of the transformant lines revealed no cloning artefacts such as mis-sense, non-sense or frameshifts mutations that might account for the lack of detectable protein. Furthermore, transcription of downstream DNA from the *atpA* promoter/5' UTR cassette has been demonstrated for *cpl-1* in Chapter three, and for numerous other genes in the Purton lab. This is would therefore suggest that the lack of detectable protein is a problem either of efficient translation or protein stability, with the former more likely as discussed below.

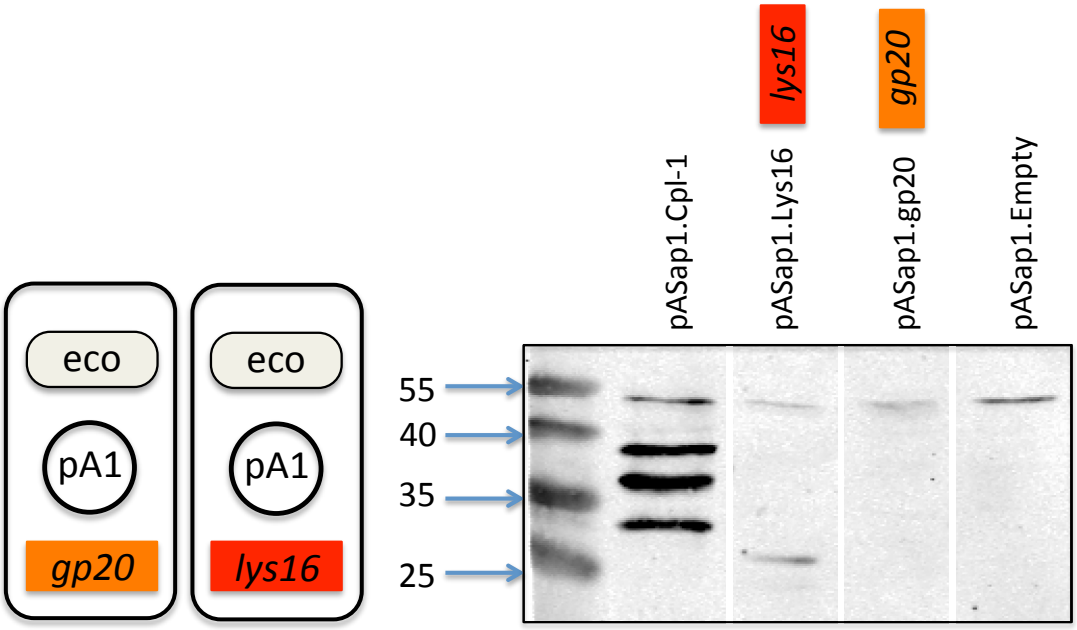


Figure 4.5 – Western blot analysis with anti-HA antibodies of *E. coli* DH5α transformed with pASap1.gp20 and pASap1.lys16 with pASap1.cpl-1 and pASap1.empty as controls.

Lys16 (29.6 kDa) is seen to be weakly accumulating, but no detectable protein is seen for Gp20 (32.4 kDa). The non-selective band at ~54 kDa indicates similar loading.

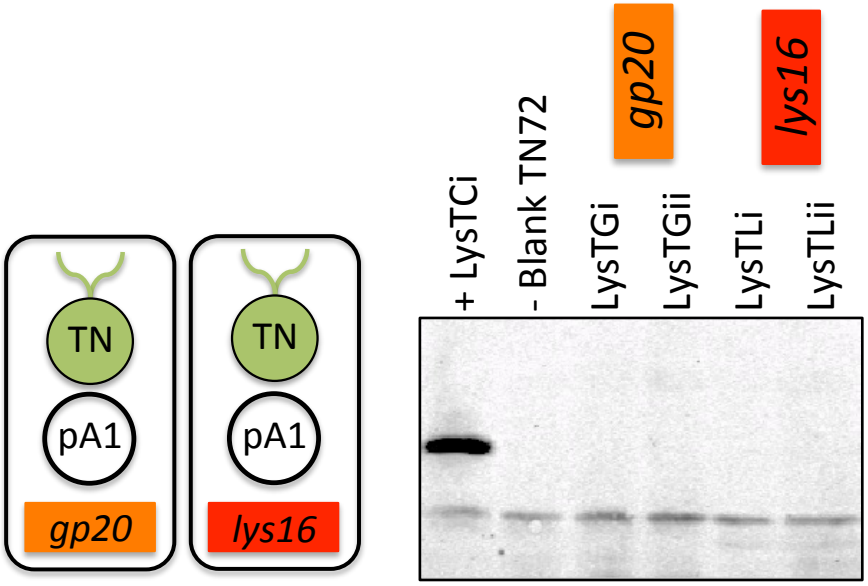


Figure 4.6 - Western blot analysis with anti-HA antibodies of *gp20* and *lys16* expressing lines of *C. reinhardtii*, LysTG and LysTL

No detectable accumulation is observed for either Gp20 (32.4 kDa) or Lys16 (29.6 kDa) despite a strong band for the *cpl-1* containing positive control. The lower non-specific band shows equal loading.

4.2.2 Novel systems for expression of non-detectable proteins

As discussed above, *lys16* and *gp20* were not seen to express to detectable levels in the *C. reinhardtii* chloroplast. Given the similarities in structure and native environment between these and the successfully expressed *cpl-1*, this was unexpected - although such failures are not uncommon, both in the Purton lab of the relevant literature. Transgene expression in the *C. reinhardtii* chloroplast is still very much in its infancy and cases of non-detectable levels of recombinant protein seem to account for about half of transgenes tested. As discussed above, there are several compelling reasons why *lys16* and *gp20* are good candidates for *C. reinhardtii* chloroplast expression and thus the direction of this project was shifted to investigate factors affecting transgene expression, with the intention of improving *lys16* and *gp20* expression to detectable levels.

4.2.2.1 ***Background to expression of non-detectable proteins***

As is often the case, negative data are rarely seen in the literature with issues of non-expressing genes being no exception. Several reports can be found however; a selection is shown in Table 4.1. These data indicate that cases of unsuccessful transgene expression are widespread, and also that absence of detection by western blot analysis does not necessarily mean that protein accumulation is not happening, albeit at extremely low levels.

Table 4.1 – A panel of *C. reinhardtii* chloroplast synthesised proteins that did not give detectable accumulation by western blot analysis

Gene product	Comments	Absolute expression	Reference
Erythropoietin	No expression detected by western blot analysis or immunoprecipitation	No expression	(Rasala *et al.*, 2010)
10FN3	Expression below lower limit for western blot analysis confirmed by immunoprecipitation. Expression detectable by western blot analysis when fused with SAA	Expression	(Rasala *et al.*, 2010)
Interferon	No expression detected by western blot analysis or immunoprecipitation	No expression	(Rasala *et al.*, 2010)
Proinsulin	Expression below lower limit for western blot analysis, but confirmed by immunoprecipitation	Expression	(Rasala *et al.*, 2010)
SHBP	No expression detected by western blot analysis	No expression	B. Mackrow and C. Economou, unpublished work
CodA (*S. cerevisiae*)	No expression detected by western blot analysis	No expression	R. Young, unpublished work
AadA	No expression detected by western blot analysis, however *aadA* phenotype observed	Expression	C. Economou, unpublished work

4.2.2.1.1 Possible caused for absence of expression

The process of gene expression, from DNA to stable protein accumulation is, needless to say, highly complex and multifaceted. As this chapter is focused on experimental investigation, the approach taken is guided by previous experimental data in regard to non-detectable proteins in the Purton lab; specifically that in each case assayed transcripts have been present even when detectable recombinant protein was not. In light of this, coupled with the well established dogma that chloroplast gene expression is largely controlled at the post-transcriptional level (Rochaix, 2001), investigations were focused on the issue of efficient translation.

4.2.2.1.2 Assumptions made in relation to stability

To narrow the field of search further, the assumption is made that both Lys16 and Gp20 are able to correctly fold into a stable form in the *C. reinhardtii* chloroplast. This supposition is based on the concept that lysins generally show a high degree of stability in prokaryotic environments, as evidenced by their exposed presence in their host bacterium's cytoplasm during the phage life cycle. The implication is that lysins are evolutionarily tailored for correct folding, and stability in their native setting. The chloroplast, as a relic of an endosymbiotic event, shares many similarities with its prokaryotic cousins and thus it is thought that such adaptations should allow for stable lysin accumulation here also. This hypothesis has at least in part been verified: Oey and colleagues have shown the lysin PlyGBS to be extremely well expressed and highly stable in the tobacco chloroplast (Oey *et al.*, 2009a), and in the preceding chapter high levels of stability have been demonstrated for Cpl-1 in crude *C. reinhardtii* extracts. Stable accumulation has also been observed for the lysin Pal in the *C. reinhardtii* chloroplast (L. Stoffels, unpublished work). Though preliminary, these results suggest that other lysins will also be stable in the *C. reinhardtii* chloroplast and hence it is assumed that the lack of detectable protein is not due to proteolytic degradation.

4.2.3 Optimisation of ribosome: transcript interactions in the translation initiation region (TIR)

Given the above assumptions that neither transcription nor proteolytic degradation is responsible for the absence of detectable recombinant lysin, translation is the most probable point for disruption of expression. Chloroplast translational control is thought mainly to be mediated at the initiation stage (Marín-Navarro *et al.*, 2007) so early attempts to bring about expression were focused on the ribosome binding site. Because endogenous expression signals are used for transgene expression in the chloroplast, and previous work (for example Cpl-1) had shown the *atpA* promoter/5' UTR combination to be functional, the most apparent unknown in the ribosome: transcript interaction complex is to be found directly downstream of the AUG, the so-called downstream box.

4.2.3.1 ***Background of downstream box interactions***

The downstream box, defined as a translation initiation enhancer in the +15 to +26 region, has long been known to be important for prokaryotic translation initiation, in some cases as important as the Shine Dalgarno sequence, if not more so (Sprengart *et al.*, 1996). Previous work conducted in *C. reinhardtii* (Kasai *et al.*, 2003) and more recently tobacco (Gray *et al.*, 2011) has shown significant increases in recombinant protein expression following downstream box optimisation, commonly via creation of endogenous: recombinant chimeras.

4.2.3.2 ***The pASap2 vector as a complete* atpA *translation initiation region (TIR)***

The pASap2 vector (Appendix j) was built by Dr Chloe Economou, and includes the first 34 codons of the endogenous *atpA* gene together with the *atpA* promoter and 5' UTR already seen in pASap1. Using this new vector, a translation fusion is created in the form of $AtpA_{34}$:GoI. Furthermore, sequence for a stromal processing peptidase (SPP) site (derived from the chloroplast imported protein PsaE) is included between the two elements to mediate auto cleavage of the chimera by SPP following its synthesis in the chloroplast. Inclusion of the complete ribosome-binding footprint from *atpA* is intended to remove any unfavourable interactions between the gene of interest and the ribosome which might be disadvantageous to translation initiation, as illustrated in Figure 4.7.

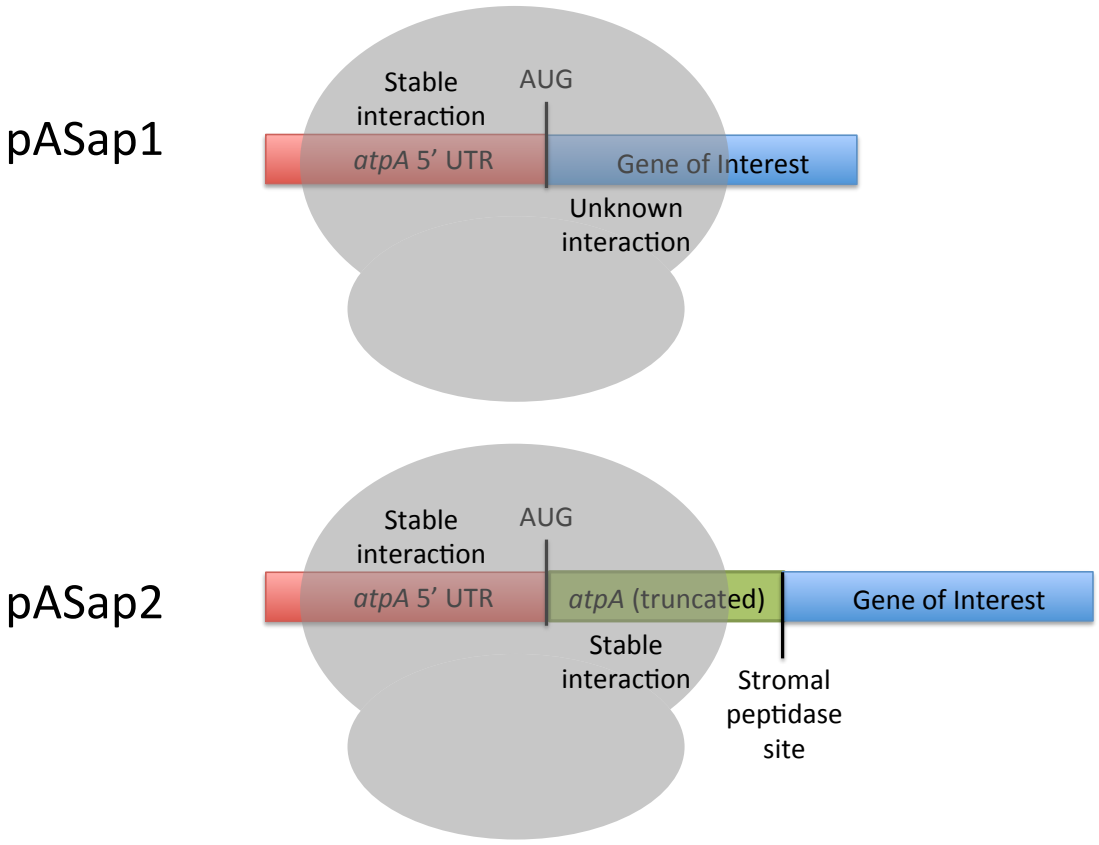


Figure 4.7 – Simplified cartoon representation of ribosome: transcript interactions in pASap1 and pASap2

pASap1 contains the *atpA* 5' UTR and as such the ribosome: transcript interactions upstream of the AUG can be considered as favourable; however, this cannot be said for those interactions downstream of the AUG. In pASap2 the first 34 codons of *atpA* are included to stabilize these interactions, included those involving the downstream box. For simplicity, only the ribosome portion of the translation initiation complex is displayed.

4.2.3.3 ***Creation of* C. reinhardtii *lines containing* lys16 *and* gp20 *utilizing the chimeric vector pASap2***

A full map of pASap2 can be found in Appendix j. Due to the similarity in vectors, including restriction endonuclease cloning sites, *gp20* and *lys16* genes were cloned into the pASap2 vector as described for pASap1 above (4.2.1.2). *E. coli* transformants were screened by PCR (Appendix k, Appendix l) and confirmed by DNA sequencing.

The plasmids pASap2.gp20 and pASap2.lys16 were used to transform the *C. reinhardtii* recipient line TN72 in the same manner as for their corresponding pASap1 constructs (4.2.1.2). Putative transformants were screened by PCR (Appendix m, Appendix n) and confirmed by DNA sequencing. Transformation frequency was further reduced relative to pASap1 transformations for *lys16* and *gp20*, yielding three transformants for pASap2.lys16, and only one for pASap2.gp20. Confirmed transgenic lines were named LysT2Gi for *gp20* and LysT2Li and LysT2Lii for *lys16*.

4.2.3.4 ***Expression of* gp20 *and* lys16 *using pASap2***

4.2.3.4.1 Expression in *E. coli*

As for pASap1 based cassettes, transgene expression was analysed in *E. coli* in addition to *C. reinhardtii*. Protein samples for *E. coli* containing pASap2.gp20 and pASap2.lys16 were prepared and analysed as for pASap1 constructs (see 4.2.1.3.1). The western blot analysis is shown in Figure 4.8. As for pASap1, no recombinant protein is detectable for Gp20. Interestingly, there is also no product seen for Lys16, indicating a reduction in accumulation relative to pASap1.lys16. It is unclear whether this is due to reduced expression or increased degradation, although in light of the *cpl-1* positive control the latter seems more likely. pASap2.cpl-1 was analysed as a vector positive control relative to pASap1.cpl-1. The full-length $AtpA_{34}$:Cpl-1 chimera is seen to accumulate to a lower level than the full length Cpl-1 generated from pASap1.cpl-1. The band for the larger truncation degradation product however is more intense, suggesting comparable levels of expression, but a less stable full length final product for the chimera. This issue is explored further in the discussion of this chapter.

4.2.3.4.1.1 Expression in the C. reinhardtii *chloroplast*

The transformed strains containing *lys16* (LysT2Li) and *gp20* (LysT2Gi) in the pASap2 cassette were analysed for recombinant protein, as for the pASap1 lines above (4.2.1.3.2). The resulting western blot (Figure 4.9) shows no expression for *gp20* or *lys16* using the pASap2 cassette. As expression was not detectable using the pASap1 construct either, it is not possible to comment on whether the $AtpA_{34}$:GoI chimera has had a positive or negative effect, but only that it was not sufficient to bring expression to detectable levels.

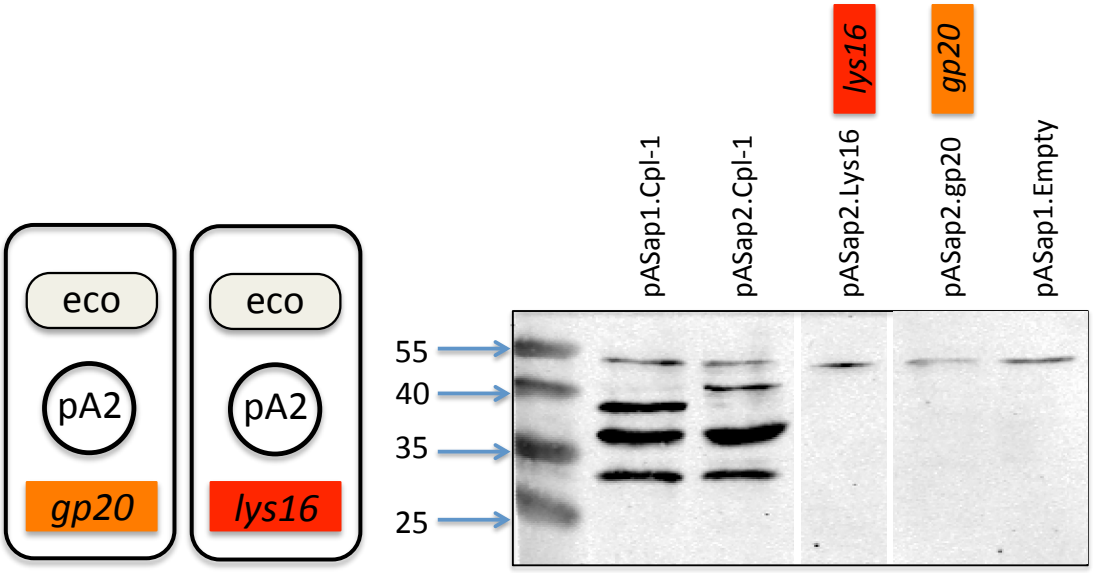


Figure 4.8 – Western blot analysis with anti-HA antibodies of *E. coli* DH5α transformed with pASap2.gp20 and pASap2.lys16 with pASap1.cpl-1, pASap2.cpl-1, and pASap1.empty as controls.

The *cpl-1* expressing positive controls show pASap2 to function as an expression vector in *E. coli*, albeit with significantly less full-length product accumulation relative to pASap1. No detectable protein is seen for either $AtpA_{34}$:Gp20 (38.7 kDa) or $AtpA_{34}$:Lys16 (36.0 kDa). Equal loading is shown by non-specific banding at 54 kDa.

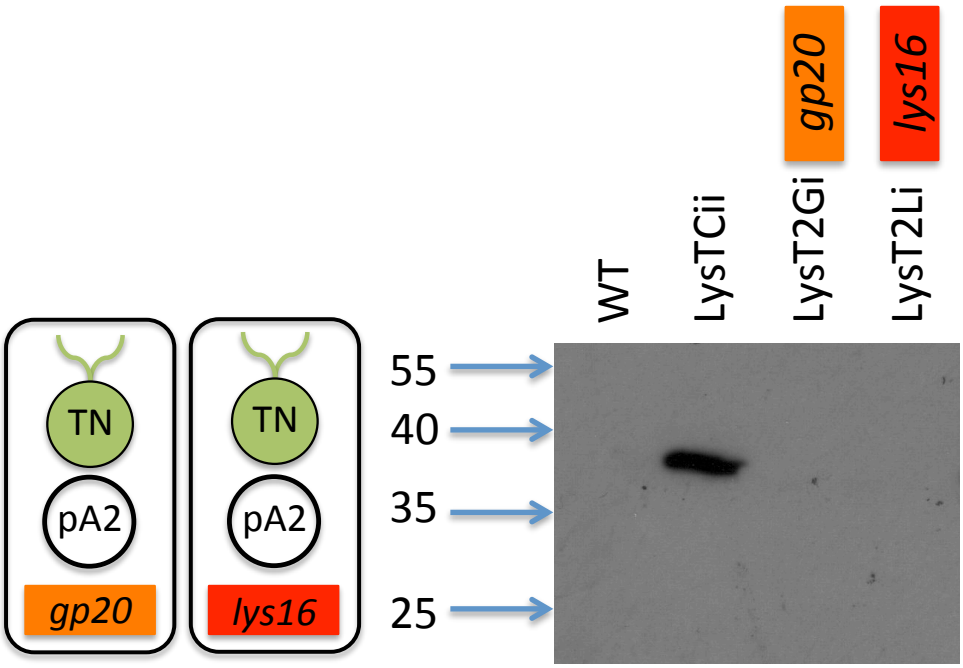


Figure 4.9 – Western blot analysis with anti-HA antibodies of *gp20* and *lys16 atpA* chimera expressing lines of *C. reinhardtii*, LysT2G and LysT2L

Neither $AtpA_{34}$:Gp20 (38.7 kDa, LysT2Gi) nor $AtpA_{34}$:Lys16 (36.0 kDa, LysT2Li) are seen to accumulate relative to a Cpl-1 positive control (40.3 kDa, LysTCii). This immunoblot was visualised by the ECL method due to higher sensitivity.

4.2.4 Bypassing initiation difficulties by use of a full fusion construct

Since the local modification of the ribosome binding site by extending the 5' *atpA* element downstream of the AUG to codon 34 failed to give detectable expression for either *lys16* or *gp20*, a different strategy was adopted. Here, the coding region of each gene is fused downstream of a full length lysin gene of proven expression to create a dual functional fusion, similar to that used by Rasala and colleagues as noted in Table 4.1. If successful this would not only bring previously undetectable proteins to detectably levels but also, if correctly folded, produce duel-specificity lysins, opening the door for further lysin customisation.

4.2.4.1 ***Advantages of full-length fusions over AtpA$_{34}$ chimeras***

Although the *atpA*$_{34}$ strategy is designed to stabilise the ribosome:transcript complex in a linear local fashion, it is unable to correct for three dimensional interactions from further downstream elements folding back towards the ribosome, or long range mRNA secondary structure interactions. By attaching an entire gene that has been shown to be expressed successfully upstream of the GoI such factors should be mitigated. *Scilicet*, by incorporating the entire coding sequence of a gene with confirmed expression it can be seen as highly likely that translation initiation at least will be successful in a fusion context. Failure to detect protein in such a system would thus represent complications at a post initiation stage.

4.2.4.2 ***Development of fusion lysins from a synthetic biology prospective***

The potential of natural lysins as novel therapeutic agents has already been discussed in this thesis, and more so in the wider literature (Fischetti, 2010; Schmelcher *et al.*, 2012). They also, however, hold great promise as a starting platform for a whole generation of customised proteins created using synthetic biology approaches. Design alterations could include modification of cell wall binding domains for more efficient detachment from the cell wall following catalysis, domain swapping for improved pharmaceutical properties, and of course domain oligomerisation for specific multi-target lysins. The creation of fusion lysins explored in this chapter can be seen as not only an attempt to raise the

expression of difficult recombinant proteins to detectable levels, but also as a first step towards such targeted broad-spectrum treatments.

4.2.4.3 ***Design, construction, and cloning of fusion constructs***

As Cpl-1 was most highly accumulating recombinant protein expressed in the Purton lab at the time, and had been shown to be a functional lysin, *cpl-1* was chosen as the upstream gene for the three fusions to be developed. The *gp20* and *lys16* genes were used as downstream test cases, with the well-expressed lysin gene, *pal* as a positive control (L. Stoffels, unpublished work). In order to allow unhindered folding and activity of the two lysin enzymes within the fusion, a 14 residue flexible linker region was designed based on a motif utilised by Mayfield and colleagues for construction of a single-chain antibody in *C. reinhardtii* (Mayfield *et al.*, 2003). A schematic of the fusion constructs can be seen in Figure 4.10.

Fusion gene constructs were built using the previously synthesized genes *cpl-1, gp20, lys16,* and *pal*. Primers were designed incorporating complementary 5' extensions encoding the linker region, and PCR conducted in two stages as shown in Figure 4.11. Stage one consisted of the addition of a downstream extension for *cpl-1*, and upstream extensions for *gp20, lys16,* and *pal*. In stage two, pairs of up- and downstream genes were mixed and treated to 10 cycles of PCR in the absence of primers. Flanking primers at each end of the newly created fusion were then added and a standard PCR reaction conducted as described above (2.2.5).

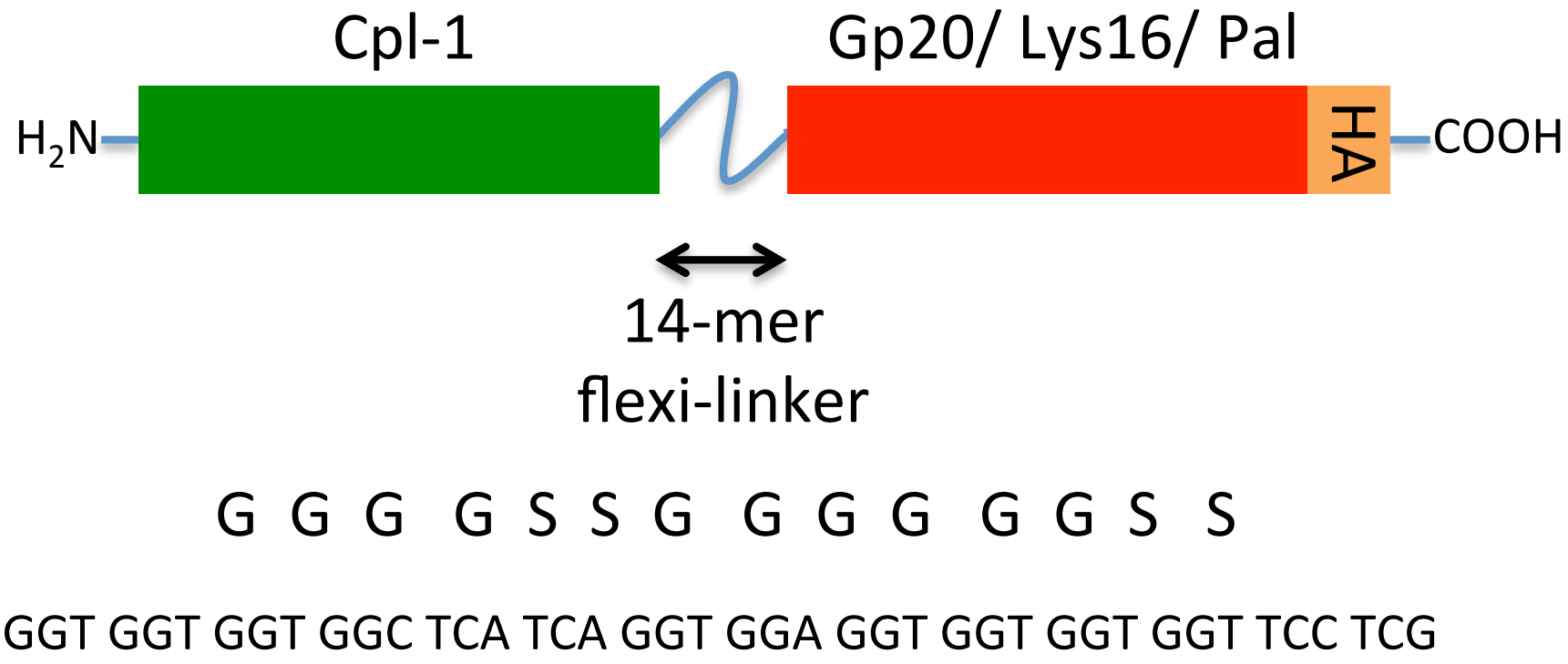


Figure 4.10 – General schematic representation of the three fusion lysins investigated.

In each of the three fusion lysins the protein of interest (PoI) is preceded by Cpl-1 with the addition of a 14-mer flexi-linker, and the deletion of the Cpl-1 HA epitope tag. The C-terminal HA tag is maintained on the PoI to allow detection by immuno-blot.

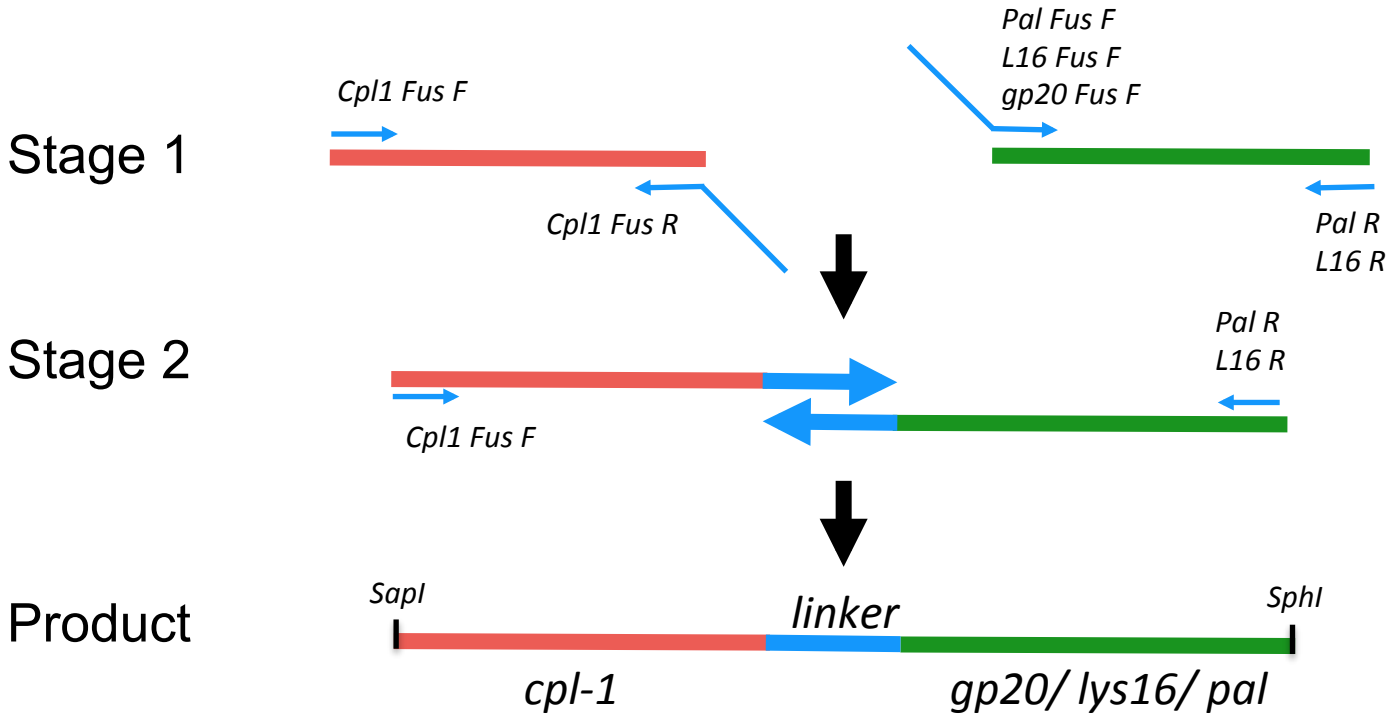


Figure 4.11 – PCR strategy for the construction of fusion lysin genes

The three fusion lysins were built utilising a two-stage PCR protocol. In stage one *cpl-1* is given a downstream extension consisting of the linker coding sequence removing the 3' *Sph*I site and HA epitope tag. The downstream GoI is given a complementary upstream extension, removing the 5' *Sap*I site. The same HA tag sequence at the 3' end of *gp20* and *lys16* allows for *L16.R* to prime both. In stage two the complementary linker regions are allowed to anneal and self-prime for ten PCR cycles before addition of flanking primers prompts amplification of the newly built fusion gene. The final product consists of *cpl-1* linked to the GoI by a flexi-linker. The entire construct contains a single HA tag sequence and unique *Sap*I and *Sph*I sites for further cloning.

4.2.4.4 ***Creation of* C. reinhardtii *transformant lines containing the fusion lysin genes***

The subsequent availability of the more highly expressing vector pSRSap1 (for details see 3.2.2.4) in the Purton lab gave rise to a dilemma in regard to the fusion lysins. pSRSap1 offered a greater chance of success due to its higher expression levels; however, this would not allow for a direct comparison with the non-detectable proteins in a stand-alone context. It was thus decided to express *cpl-1:pal* and *cpl-1:lys16* in pSRSap1 and *cpl-1:gp20* in pASap1. As pSRSap1 contains the same cloning sites as pASap1, the same strategy was employed as for pASap1 and pASap2. Confirmation of insertion of the fusion constructs into the expression vectors is presented in Appendix o and Appendix p.

The three plasmids pSRSap1.cpl-1:lys16, pSRSap1.cpl-1:pal, and pASap1.cpl-1:gp20 were used to transform the *C. reinhardtii* recipient line TN72 as described previously (4.2.1.2). Transformation yielded low numbers of colonies, but those screened by PCR (Appendix q, Appendix r) and sequencing were confirmed as true transformants. These fusion lines were named LysTC:P-SR, LysTC:L-SR and LysTC:G for pSRSap1.cpl-1:pal, pSRSap1.cpl-1:lys16, and pASap1.cpl-1:gp20, respectively.

4.2.4.5 ***Expression of fusion lysins***

4.2.4.5.1 Expression in *E. coli*

As with the single lysin constructs, transgene expression was analysed first in *E. coli*, to confirm construct assembly and to provide a comparison to *C. reinhardtii* expression. *E. coli* protein extracts were produced as previously described (4.2.1.3.1) and analysed by western blot analysis with anti-HA antibodies as shown in Figure 4.12 and Figure 4.13. All three fusion constructs show the expression of HA tagged protein; however, in each case the protein is at least partially fragmented. *cpl-1:lys16* and *cpl-1:pal* both show significantly increased expression relative to their non-fusion counterparts as expressed in pASap1. This cannot be accredited to the fusion construct however, due to the use of the *psaA* promoter/ 5' UTR. It has been previously demonstrated (Ninlayarn, 2012, submitted) that the *psaA* promoter is considerably more active than that of *atpA* in *E. coli* due to a

more defined Shine Dalgarno sequence. *cpl-1:gp20* on the other hand is expressed under the *atpA* promoter/ 5' UTR so a direct comparison can be made.

Only the fusion positive control *cpl-1:pal* shows significant levels of full length fusion protein relative to fragmentation products. *Cpl-1:gp20* appears to show full length product, although at a much lower concentration than its smaller bands. *Cpl-1:lys16* is unusual in that it shows no full length fusion protein, but unique amongst lysins expressed in *E. coli* (with the exception of singlet Lys16) gives a remarkably clean blot with a single band. Only a very weak band is seen for the full length Cpl-1:Gp20 fusion, but the truncated products (incorporating full length Gp20) are shown to be at a similar intensity to the Cpl-1 positive control. This represents a significant increase in recombinant protein accumulation in *E. coli* and implies that translation initiation may well have been the limiting factor in singlet *gp20* expression.

From these data it is unclear whether the fragments observed are due to proteolytic degradation or cryptic internal initiation sites. It should be noted however, that the majority of truncation products seen are larger than the C-terminal protein alone, implying fragmentation is occurring in the Cpl-1 portion of the fusion. This topic is discussed further below (4.2.4.6).

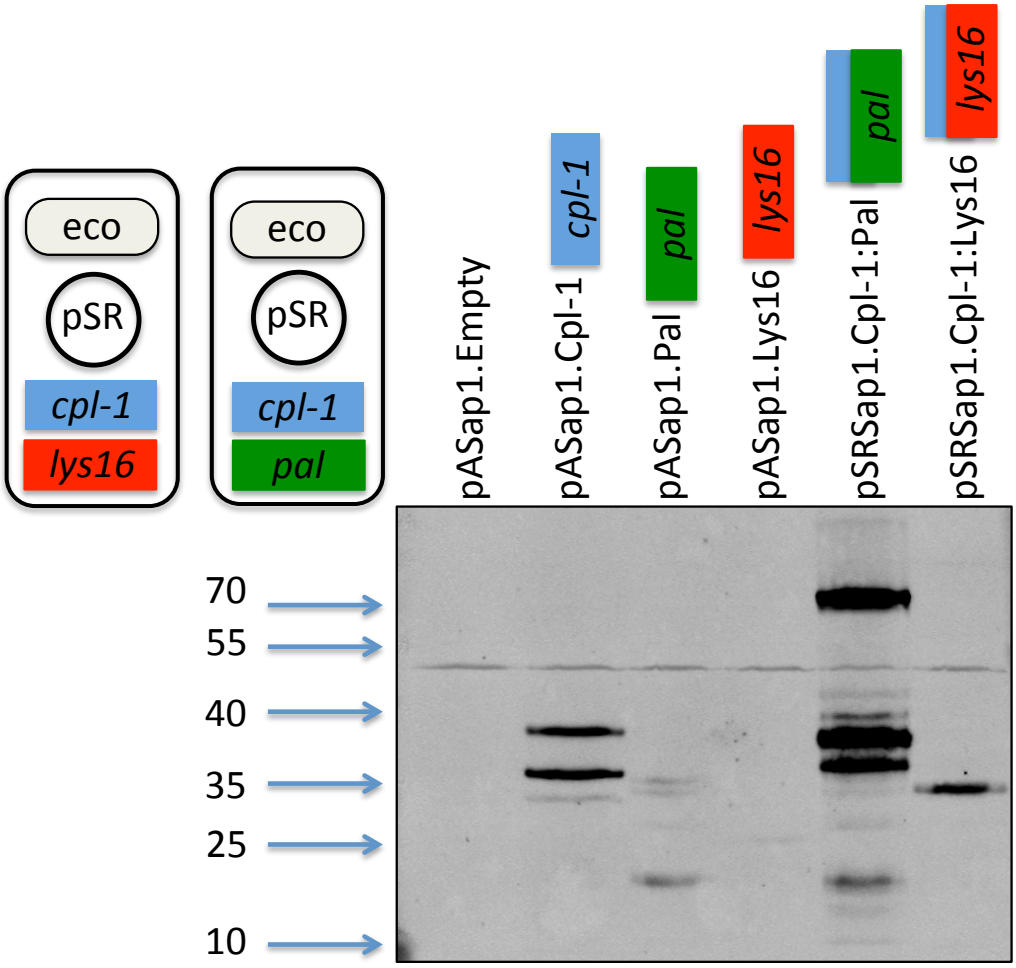


Figure 4.12 – Western blot analysis with anti-HA antibodies of *E. coli* DH5α transformed with pSRSap1.cpl-1:pal and pSRSap1.cpl-1:lys16 with various controls

This blot shows both fusion lysins to be expressed, but while the full length fusion is seen for Cpl-1:Pal (75.6 kDa), only a truncated product is seen for Cpl-1:Lys16 (69.7 kDa). Cpl-1 (40.3 kDa), Pal (35.5 kDa), and Lys16 (29.6 kDa) expressed individually are shown alongside for comparison. The fusions are shown to accumulate to greater levels than their component lysins; however, this may due to the different vectors used.

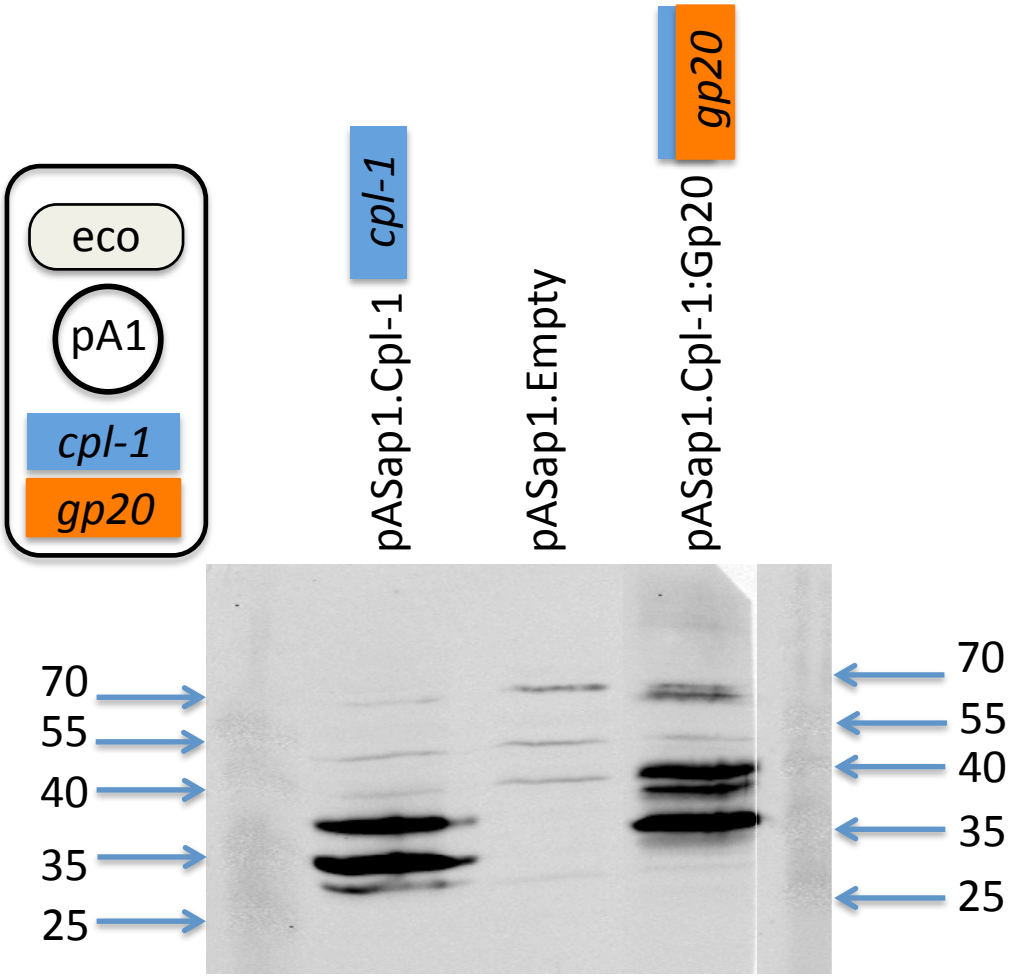


Figure 4.13 – Western blot analysis with anti-HA antibodies of *E. coli* DH5α transformed with pASap1.cpl-1:gp20

The Cpl-1:Gp20 (72.5 kDa) fusion is shown to accumulate to a comparable level to that seen for Cpl-1 (40.3 kDa) in *E. coli*, but very little of the full-length fusion is seen. Despite this absence, these data show the creation of a fusion protein to greatly increase the accumulation of a previously un-detectable recombinant product. As any product was anticipated to be at low concentration, a secondary antibody concentration of 1:10,000 was used (twice usual concentration) giving rise to the extra non-specific bands seen at 70, 40, and 30 kDa.

4.2.4.5.2 Expression in the *C. reinhardtii* chloroplast

C. reinhardtii protein extracts were produced as previously described (4.2.1.3.2), with the exception that *cpl-1:gp20* containing lines (LysTC:G) samples were prepared to double standard concentration. Western blot analyses with anti-HA antibodies are shown in Figure 4.14 and Figure 4.15. In contrast to the results from *E. coli* expression, only the fusion positive control *cpl-1:pal* gave detectable levels of protein accumulation. The absence of degradation products for this fusion implies stability, and thus it is likely that correct folding is also taking place. The fact that no recombinant protein is seen at all the either *cpl-1:lys16* or *cpl-1:gp20* is informative for several reasons. It is known that the promoter/ 5' UTR/ downstream box is active and capable of expression from the accumulation of Cpl-1 and Cpl-1:Pal. It is also evident that Cpl-1 fusion proteins are not inherently unstable in the *C. reinhardtii* chloroplast as seen from the accumulation of the Cpl-1:Pal fusion. It is also highly likely that the Lys16 and Gp20 lysins are stable in this environment as discussed previously. Taken together, these factors indicate a failure in translation elongation within the *lys16* and *gp20* regions of the two fusions to be the causative factor in non- expression in these cases. A major factor in progression of the elongation complex along a transcript is known to be the codon adaptation, and thus this was investigated next.

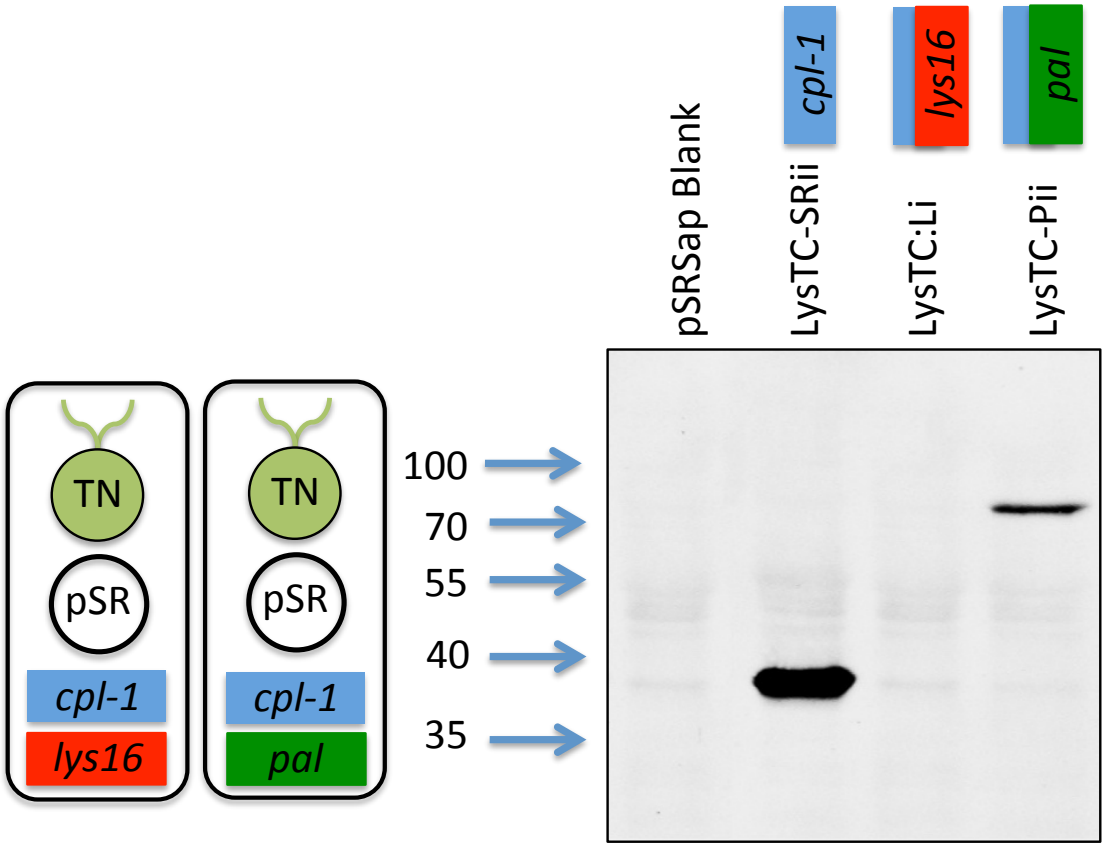


Figure 4.14 – Western blot analysis with anti-HA antibodies of *cpl-1:lys16* and *cpl-1:pal* expressing lines of *C. reinhardtii*, LysTC:Li and LysTC:Pii

No detectable protein accumulation is seen from the Cpl-1:Lys16 fusion (69.7 kDa, LysTC:Li). Full length product is, however, observed for Cpl-1:Pal (75.6 kDa, LysTC:Pii), indicating that the lack of expression seen for *cpl-1:lys16* is not a general failure of fusion design. *cpl-1:pal* is seen to express to a significantly lower level that *cpl-1* under the same promoter (40.3 kDa, LysTC-SRii).

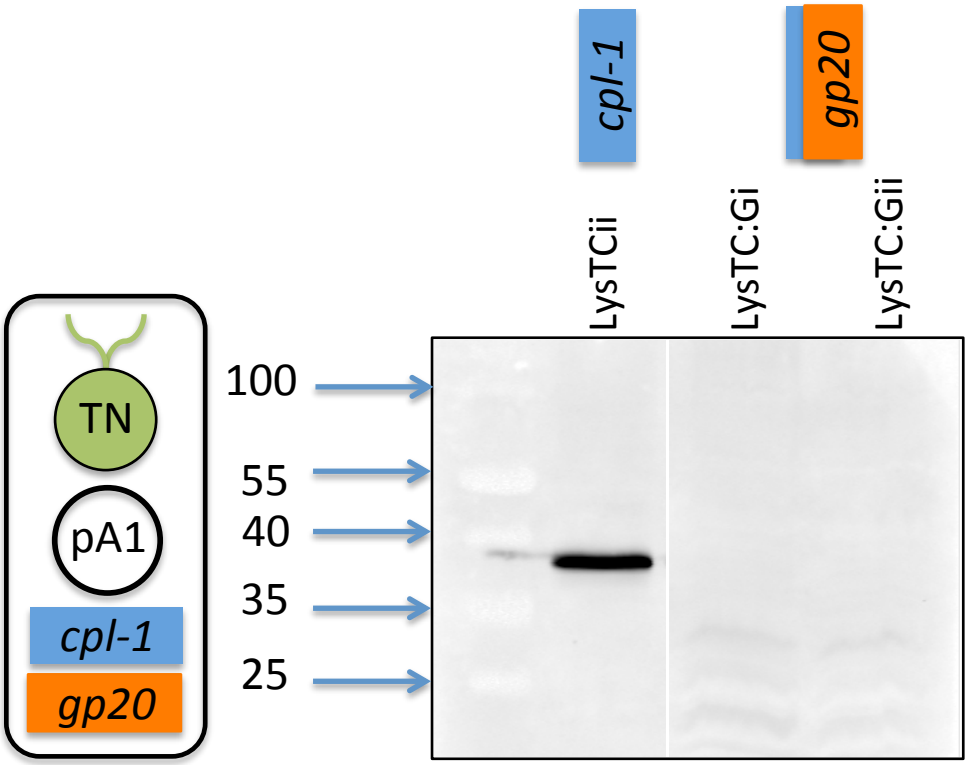


Figure 4.15 - Western blot analysis with anti-HA antibodies of *cpl-1:gp20* expressing lines of *C. reinhardtii*, LysTC:Gi and LysTC:Gii

No detectable accumulation is seen for Cpl-1:Gp20 (72.5 kDa, LysTC:Gi and ii), suggesting potential elongation abortion signals in the *gp20* transcript. *LysTC:Gi* and *LysTC:Gi* were both loaded at double concentration relative to the *LysTCii* .

4.2.4.6 ***Analysis of fusion protein degradation products***

It was noted that expression of fusion lysins in *E. coli* resulted in protein fragmentation in all cases, and also that the majority of fragments seen were within the Cpl-1 portion of the fusion. As with the Cpl-1 fragments observed for Cpl-1 above (3.2.2.1), it is unclear whether these are due to proteolytic cleavage or internal translation initiation events. To investigate if regions of fragmentation were clustered together in the three expressed fusions, predicted fragment sizes were aligned. N-terminal alignment was used to standardize for the Cpl-1 region irrespective of C-terminal domain size.

Figure 4.16 shows a scaled N-terminal alignment with approximate size values for each fragment. The shaded box indicates a region of high cleavage or internal translation initiation points. Primary sequence analysis of Cpl-1 shows a high concentration of methionines in approximately the same region as the fragmentation events are occurring (Figure 4.17). Despite this, given the range of fragment sizes seen for the different fusions it seems more probable that they are the result of proteolytic cleavage. The products of such cleavage events are then stabilised to a greater of lesser degree by the different C-terminal protein in each case.

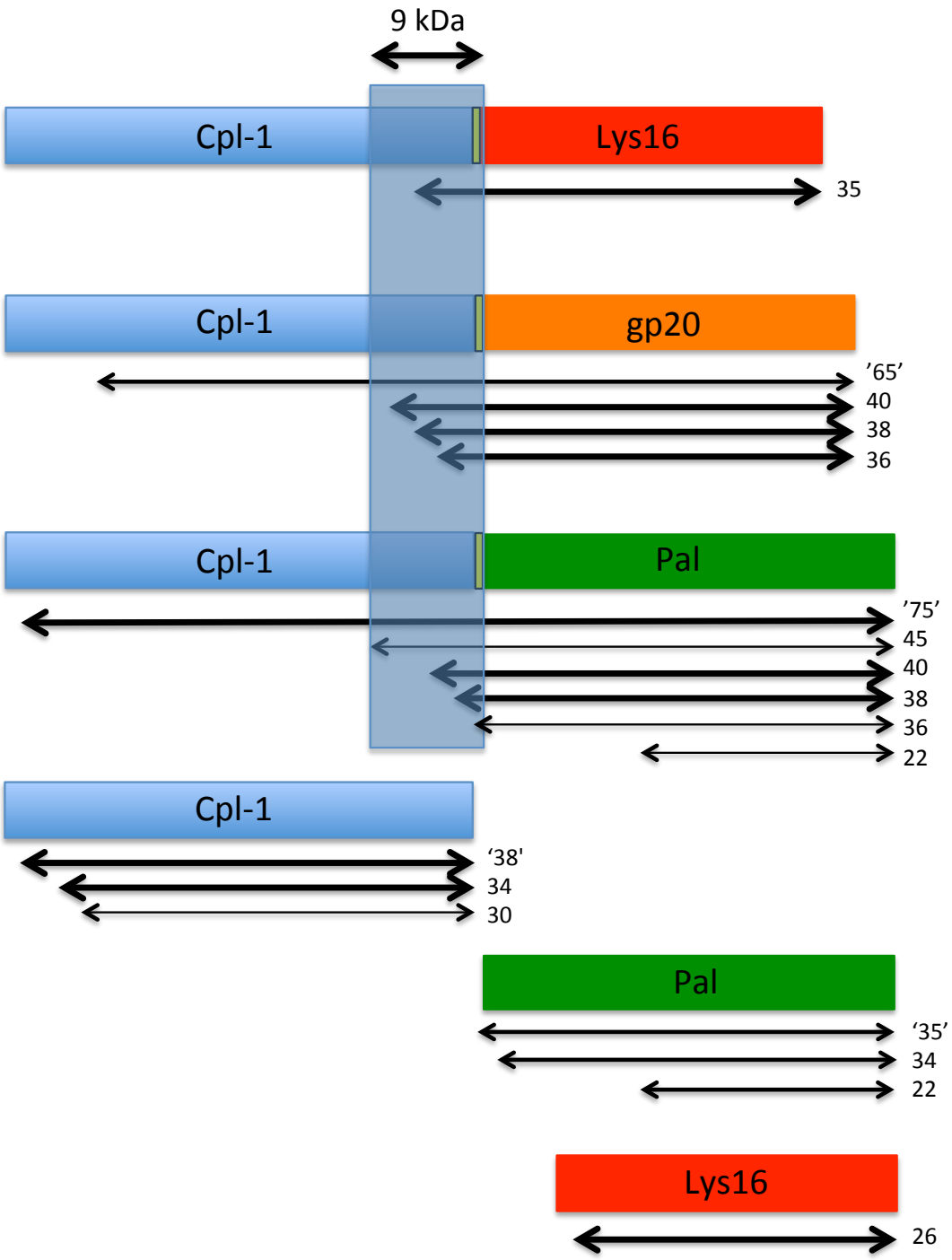


Figure 4.16 – To-scale N-terminal alignment of fusion protein fragmentation products in *E. coli* DH5α

Size values for each fragment were approximated from the western blot analyses shown above. Alignment the fusion proteins by their Cpl-1 domains shows the majority of fragments to have N-termini in the final 9 kDa region of Cpl-1. Whether this is due to high protease activity or cryptic ribosome binding events is unclear. Line thickness indicates abundance of fusion product.

Figure 4.17 – Primary sequence analysis of Cpl-1 showing methionines in the region of most fusion fragmentations

The primary amino acid sequence of Cpl-1 shows multiple methionine residues in the region of predicted fragmentation revealed by Figure 4.16, indicating possible cryptic internal translation initiation sites.

4.2.4.7 ***Purification of the* C. reinhardtii *derived Cpl-1:Pal fusion lysin***

As the only fusion lysin successfully expressed in the *C. reinhardtii* chloroplast, the Cpl-1:Pal fusion was purified following the same protocol as employed for Cpl-1 in Chapter three. This protocol was suitable in two respects: firstly, the Cpl-1 domain of the fusion would interact with the ion exchange column in the same way as for Cpl-1 alone (assuming correct folding of the fusion protein), and secondly, Pal contains a very similar choline binding domain to that seen in Cpl-1, making the purification protocol appropriate for both enzymes.

Elution fractions were analysed by direct spotting of samples onto nitrocellulose followed by immunoblot with anti-HA antibodies (Figure 4.18) as described (2.4.3.1). Fractions 4, 5, 6, and 7 were shown to contain the HA-tagged protein and were retained. The successful recovery of the Cpl-1:Pal fusion lysin by this method indicates the correct folding of at least one of the choline binding domains in the fusion molecule.

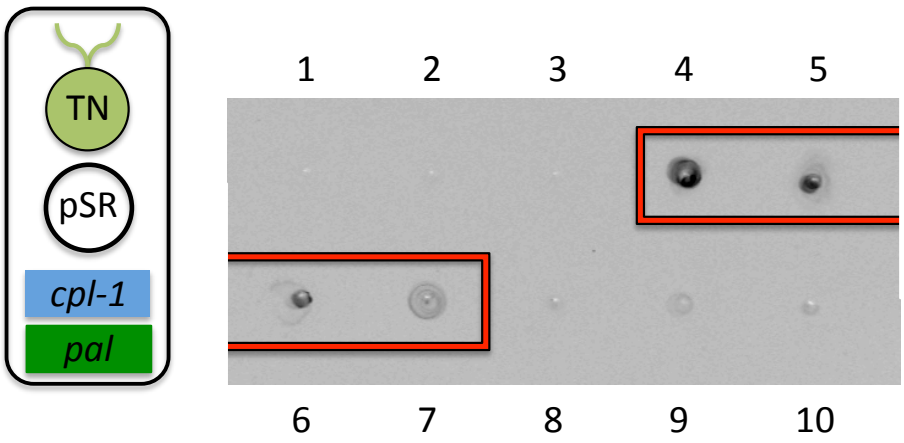


Figure 4.18 – Dot-blot analysis with anti-HA antibodies of Cpl-1:Pal fusion lysin elution fractions from DEAE ion exchange purification

Fractions 1-3 show no detectable HA-tagged protein, whereas fractions 4-7 show the highest concentrations, with levels tailing off in the later fractions. The red box indicates fractions pooled for concentration and further analysis. The successful recovery of HA-tagged protein by this method indicates correct folding of at least one choline binding domain in the fusion protein.

4.2.4.8 ***Functional evaluation of fusion lysins***

Functional analysis of fusion lysins expressed in both *E. coli* and *C. reinhardtii* was conducted as for Cpl-1 in the preceding chapter: by spectrographic liquid clearance assay of the target bacterium. Target bacterial suspensions were prepared and treated with crude cell extracts as described (2.6). Lysis of bacterial cells was observed as a function of optical density at 600 nm.

4.2.4.8.1 Cpl-1:Pal Fusion

4.2.4.8.1.1 Activity of *E. coli*-produced fusion protein

The Cpl-1:Pal fusion as synthesised by *E. coli* was shown to be highly active, with rapid clearance observed within minutes of application. *S. pneumoniae* samples were treated with 50, 100, and 200 µl of lysin preparation, and a dose-dependent drop in optical density observed relative to a control sample for *E. coli* containing the empty pSRSap1 plasmid (Chart 4.1). Owing to the highly fragmented nature of the Cpl-1:Pal fusion in *E. coli* it was not possible to say at this stage if activity was due to the full length fusion lysin, or the fragmented forms thereof. It was noted that optical density was plateauing before total clearance was achieved. In order to investigate whether this was due to enzyme or substrate depletion a preliminary experiment was conducted. Upon reaching a plateau either additional enzyme or substrate (*i.e.* bacteria) was added. Chart 4.2 shows a continuation of activity only after addition of further substrate suggesting that the plateau effect is due to the depletion of lys-able bacterial cells.

4.2.4.8.1.2 Activity of *C. reinhardtii*-produced fusion protein

In the *C. reinhardtii* chloroplast Cpl-1:Pal is synthesised as a full length fusion lysin with no evidence of incomplete fragments. This allows for analysis of the activity of the fusion enzyme without the complications seen in *E. coli*. Crude protein extracts were prepared as for Cpl-1 activity assays described in section (2.4.1.2.2). Activity of the Cpl-1:Pal fusion was assayed alongside a crude Cpl-1 preparation as a positive control. Chart 4.3 shows no detectable activity of the full-length fusion lysin relative to the pSRSap.empty blank sample. It is thought that the absence of activity is due to steric hindrance between the two lysin modules.

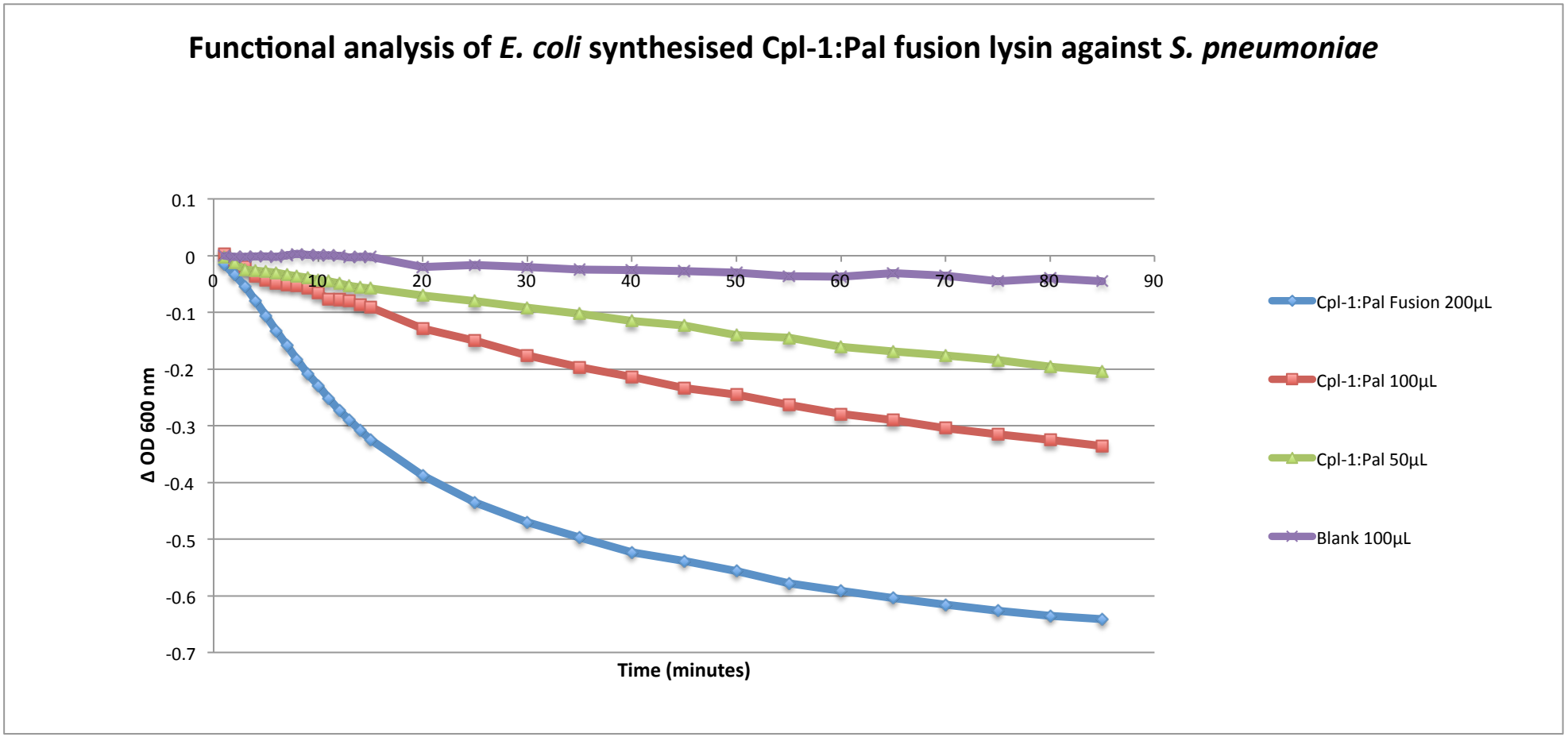
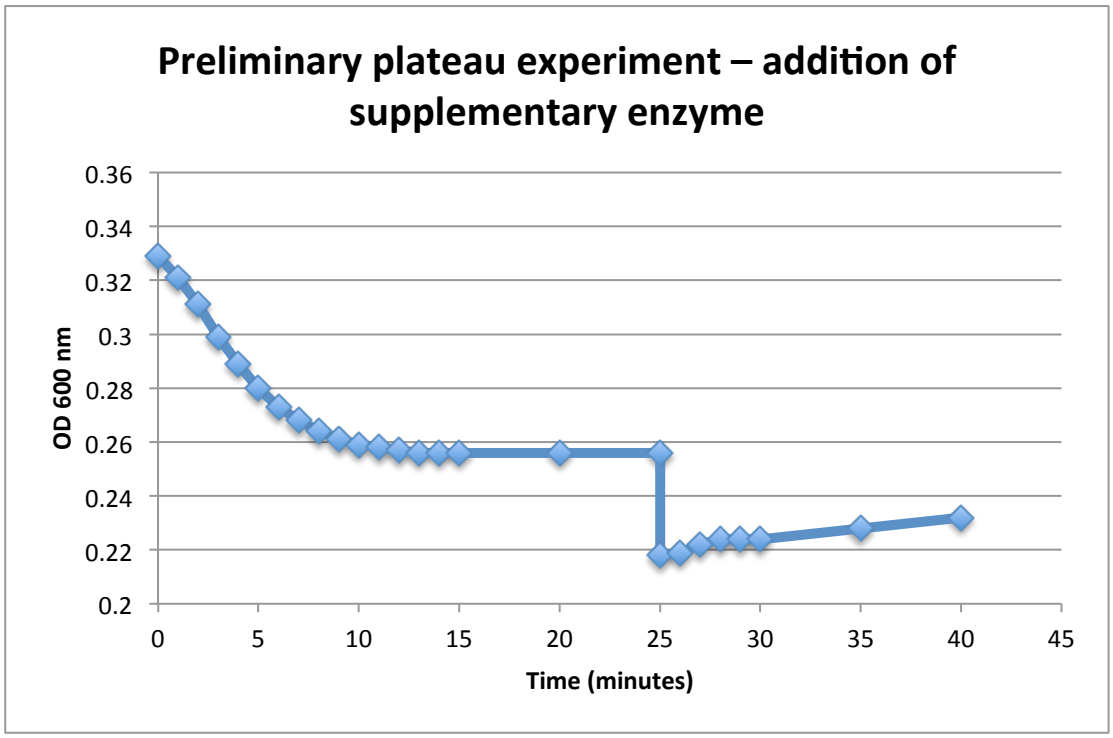


Chart 4.1 – Functional analysis of *E. coli* synthesised Cpl-1:Pal fusion lysin

Crude protein preparations of *E. coli* expressing the *cpl-1:pal* cassette show considerable lytic activity against a suspension of *S. pneumoniae* in PBS incubated at 37 °C. The degree of activity is also shown to be dependent on the amount of protein preparation added.



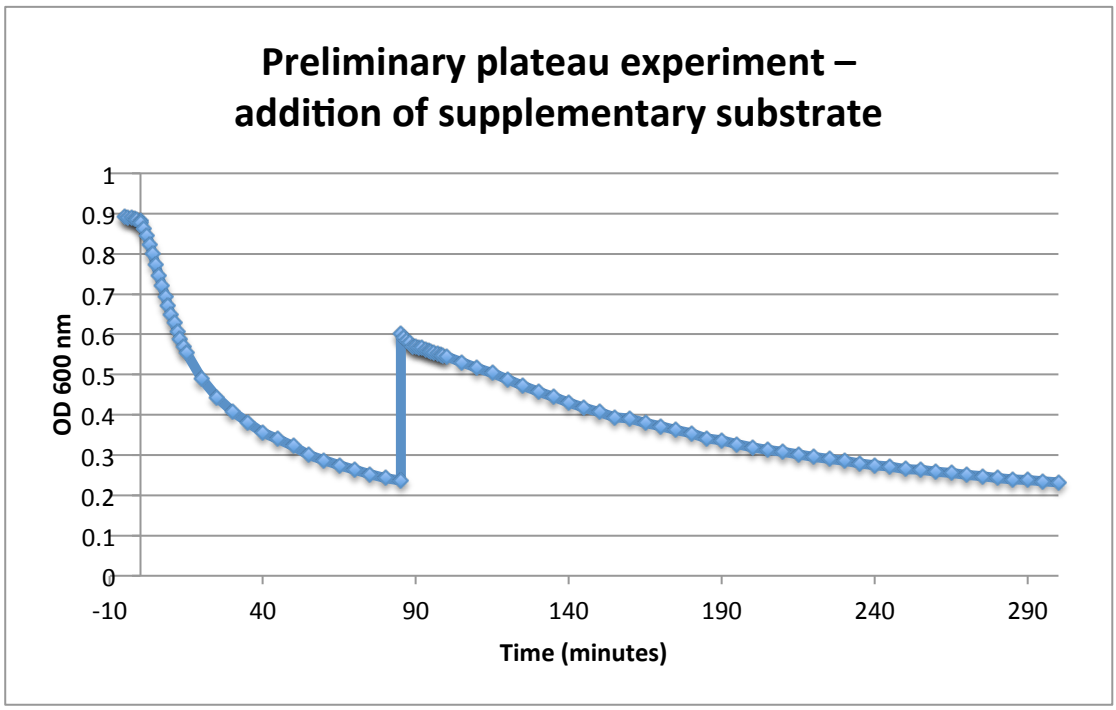


Chart 4.2 – Preliminary investigations into plateau effect observed in activity assays

Early investigations into Cpl-1:Pal clearance assays showed a plateau effect. In the upper panel a further 100 µl of crude protein preparation was added on reaching a plateau, whereas in the lower plane an additional 1 ml of *S. pneumoniae* suspension was added. It can be seen that only addition of supplementary substrate gave a continuation of activity.

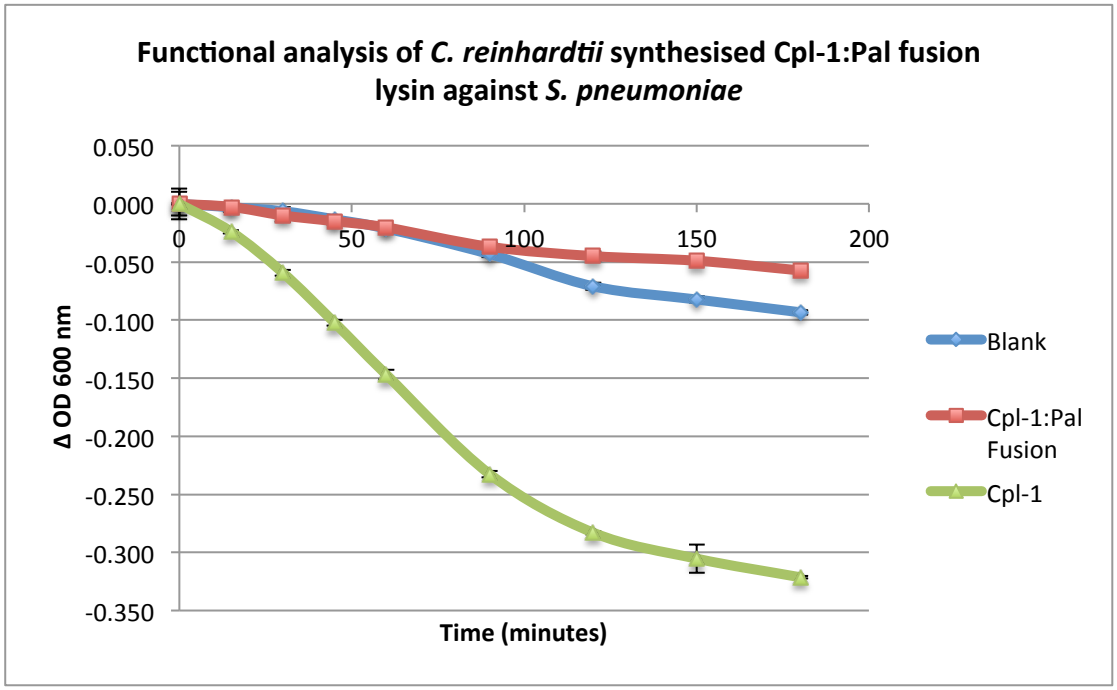


Chart 4.3 – Functional analysis of *C. reinhardtii* synthesised Cpl-1:Pal fusion lysin

No lytic activity is seen for the Cpl-1:Pal fusion when synthesised in *C. reinhardtii*, implying that activity seen for the fusion expressed in *E. coli* may have been due to non-fusion truncation fragments.

4.2.4.8.2 Cpl-1:Lys16 and Cpl-1:Gp20 Fusions

The western blot analysis of the Cpl-1:Lys16 fusion as expressed by *E. coli* (Figure 4.12) indicated that no full length fusion is present. Due to the size of the truncation it can be concluded that a considerable portion of Cpl-1 (including all of the catalytic domain) is not present and thus activity against *S. pneumoniae* was not assayed. Instead, activity of the C-terminal Lys16 domain was assayed by spectrographic clearance assay against *S. aureus* with *E. coli* derived Lys16 as a positive control and pASap1.blank in *E. coli* as a negative control.

The clearance assay in Chart 4.4 shows no significant drop in optical density relative to the positive and negative controls, indicating that, although the Lys16 module is present in a higher concentration for the fusion construct, it is not active. This is likely due to the truncated region of Cpl-1 interfering with either global folding or specific substrate interactions, possibly in a similar manner to that of *C. reinhardtii* derived Cpl-1:Pal fusion lysin.

Activity assays of the Cpl-1:Gp20 fusion were conducted by Laura Stoffels in the Purton lab. Preliminary experiments show no activity of the fusion against the Gp20 target, *P. acnes* (data not shown). This also suggests complications due to the presence of two lysins in close proximity, implying that the flexi-linker may not be sufficient to allow free movement and thus activity of the individual lysin modules.

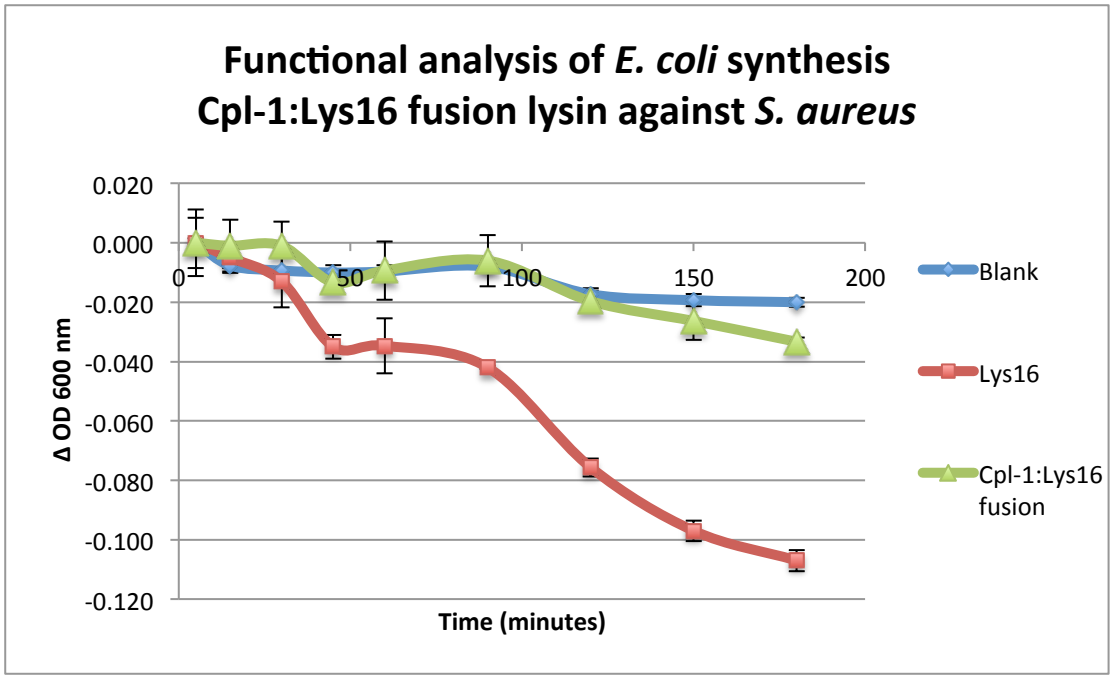


Chart 4.4 – Functional analysis of *E. coli* synthesis Cpl-1:Lys16 fusion lysin against *S. aureus*

From this experiment it is not clear if the Cpl-1:Lys16 fusion has any detectable activity relative to a pSRSap1.empty negative control (Blank). Any slight activity that is present is considerably lower than that seen for the Lys16 protein alone, despite the fusion being present at a far higher concentration as shown by western blot analysis (Figure 4.12).

4.2.5 Redesign of the *gp20* coding sequence to prevent ribosome stalling

Following the lack of detectable expression for *cpl-1:lys16* and *cpl-1:gp20* in *C. reinhardtii*, and the conclusions drawn as to the possible cause, investigations were shifted onto the next stage in gene expression: translation elongation. As with translation initiation, elongation is affected by a multitude of factors. For this section it was decided to investigate a feature generally accepted to be of importance, but neglected in terms of rigorous study in the *C. reinhardtii* chloroplast; that is the optimization of codon use.

4.2.5.1 *Novel codon optimisation techniques in the Purton lab*

A detailed overview of codon- and codon pair use is presented in Chapter five. As a very brief introduction, codon optimisation is the process by which transgenes are adapted to display similar codon preferences as seen in the host organism. During the course of the investigations discussed in Chapter five several flaws were identified with the conventional methods of codon optimisation in the *C. reinhardtii* chloroplast. To remedy such issues a new software platform, the Codon Usage Optimizer, was developed in a collaborative project with Khai Kong Jien, an undergraduate project student in the group. By utilisation of this new software the *gp20* gene was entirely redesigned, resynthesized, and cloned into pASap1.

4.2.5.2 *Redesign of the* gp20 *coding sequence*

The *gp20* coding sequence was redesigned using the CUO Moptomiser subroutine, which allows for semi-automated codon- and codon pair optimisation. A custom-made codon usage table based on a subset of highly expressing *C. reinhardtii* chloroplast genes, *C. reinhardtii chloroplast handpick*, was used to calibrate the optimiser. In total, 127 silent modifications were made, giving the newly designed gene 85.7 % DNA sequence identity to the GeneArt optimised *gp20*. The difference in codon usage is illustrated graphically in Chart 4.5. Panel A shows a 6 point moving average of codon- and codon pair CAI scores for *gp20* (calculated by the CUO program) as optimised by the GeneArt optimiser running the Kazusa codon usage table. Panel B shows the same gene after optimisation with the CUO and the newly generated high expressing codon usage table, *C. reinhardtii chloroplast handpick*. Inevitably, a gene optimised then analysed by the same program will

show a near perfect codon usage. However, given as the codon preferences used are those seen in the most highly expressing endogenous genes, Chart 4.5B is in fact indirectly displaying the degree of similarity between the newly designed *gp20* and such native genes. A 6-point moving average was used to make the data more visually accessible, and to highlight regions of particularly bad codon or codon pair use as opposed to isolated point events. Optimisation was conducted under the constraint of avoiding all *Sap*I and *Sph*I sites so to facilitate cloning into the pASap1 vector.

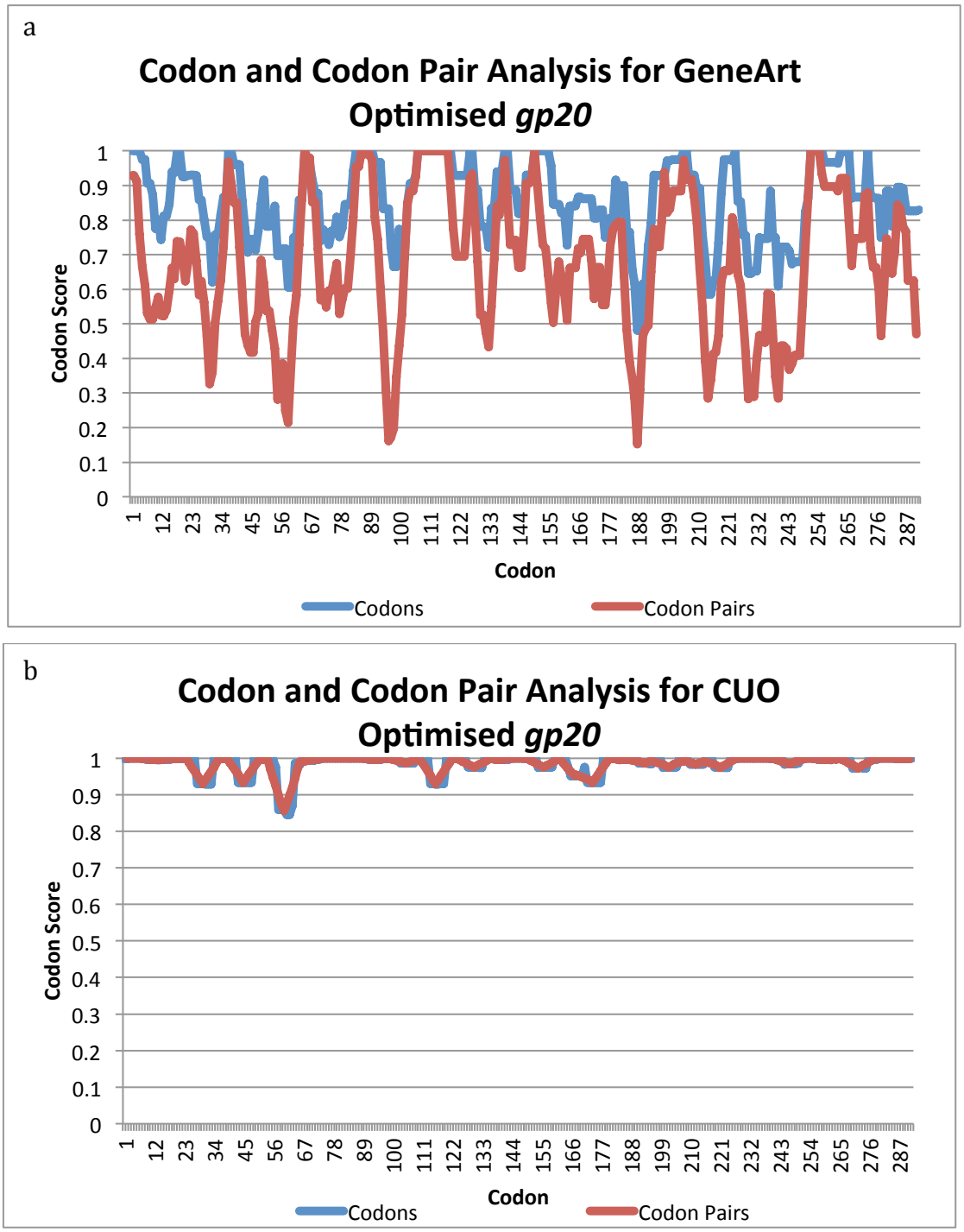


Chart 4.5 – Codon- and codon pair analyses for GeneArt and CUO optimised versions of the *gp20* coding sequence

Codon- and codon pair use is shown on a 0-1 scale, with 1 being ideal codon usage according to the codon bias used for the analysis, in this case *C. reinhardtii chloroplast handpick*. Data is presented as a 6-point moving average so to highlight regions of poor usage and mitigate isolated incidents. Panel A shows codon use, and particularly codon pairing, for the GeneArt optimised gene to be very poor relative to panel B, the CUO optimised gene.

4.2.5.3 ***Creation of* C. reinhardtii *lines containing* gp20rd**

The newly designed gene, termed *gp20rd*, was synthesised by GeneArt and cloning was conducted as previously described (4.2.1.2) employing the pASap1 vector so to provide a true comparison with the GeneArt optimised *gp20* containing line LysTG. Insertion of the target gene into pASap1 was confirmed by endonuclease digestion (Appendix s), and DNA sequencing.

Transformation was conducted by the glass bead method as previously described (4.2.1.2). Due to time constraints and limited DNA yields, only two transformation reactions were conducted, resulting in only a single transformant colony. The transformant was confirmed to be correct by PCR and DNA sequencing, and was termed LysTGrd.

4.2.5.4 ***Expression of* gp20rd**

4.2.5.4.1 Expression in *E. coli*

E. coli protein extracts were produced as previously described (4.2.1.3.1), and analysed by western blot analysis with anti-HA antibodies as shown in Figure 4.19. As with Figure 4.15, double the standard secondary antibody was used so to lower the detection threshold; however, no Gp20 is observed at the expected size of 32 kDa.

4.2.5.4.2 Expression in the *C. reinhardtii* chloroplast

The transformed strain LysTGrd, was grown to late log phase in TAP medium. Harvested cells were equalised to double normal cell concentration and prepared for western blotting as described (2.4.1). Samples were analysed by western blot analysis with anti-HA antibodies visualised using the LiCor odyssey system. Figure 4.20 shows no detectable expression for the redesigned *gp20rd*, indicating that the previous absence of *gp20* expression was not simply the result of poor codon optimisation.

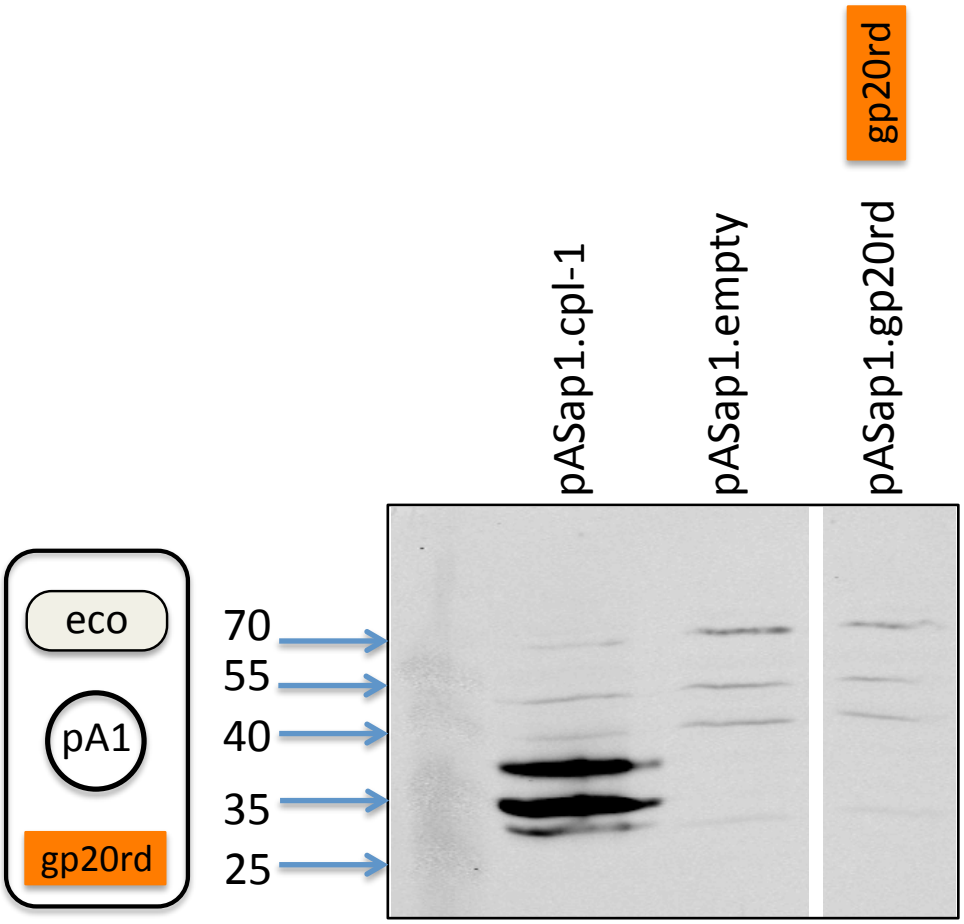


Figure 4.19 – Western blot analysis with anti-HA antibodies of *E. coli* DH5α transformed with pASap1.gp20rd shows no Gp20 accumulation

In this western blot double the standard secondary antibody was used (1:10,000) so to lower the detection threshold; however, no Gp20 is observed at the expected size of 32.4 kDa.

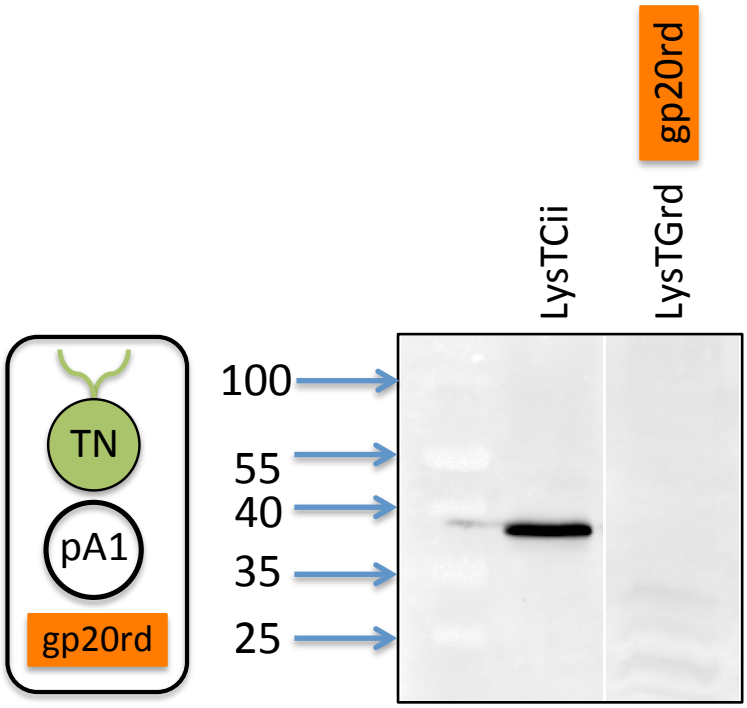


Figure 4.20 – Western blot analysis with anti-HA antibodies of the *gp20rd* expressing line of *C. reinhardtii*, LysTGrd shows no Gp20 accumulation

LysTGrd was loaded at double normal equalised concentration. No detectable Gp20 is observed at 32.4 kDa.

4.3 Discussion

Following the successful production of Cpl-1 in the *C. reinhardtii* chloroplast, two new lysins were selected. The first, Lys16, from the *S. aureus* phage P68, had been previously synthesised and partially characterised in *E. coli*, but this platform yielded issues with solubility. *C. reinhardtii* chloroplast expression held the potential to mediate such problems, owing to the presence of eukaryotic chaperones among other factors. The second, Gp20, a putative lysin from the *P. acnes* phage PA6, holds promise as a cosmetic product that would be particularly suited to expression in a Generally Recognised As Safe (GRAS) organism. Both lysin genes were cloned into the *C. reinhardtii* chloroplast as for *cpl-1* in the preceding chapter, however neither yielded any detectable protein as assayed by anti-HA western blot. Three novel approaches were investigated to bring about expression of these genes, exploring issues of local translation initiation, longer range factors affecting translation initiation, and the optimisation of codon use to improve translation elongation.

4.3.1 Manipulation of the downstream box: effects of pASap2 on expressing and non-expressing proteins

Previous optimisation of translation initiation in the *C. reinhardtii* chloroplast has focused almost exclusively on sequence upstream of the AUG (i.e. the 5' UTR) and presumed that sequences downstream of the AUG (i.e. within the coding region) play no part. The use of the chimeric expression vector pASap2 allows faithful reproduction of the endogenous translation initiation region from *atpA* both up and downstream of the AUG, conserving the so-called downstream box.

In the case of *lys16* and *gp20* this did not result in detectable levels of expression using either the *E. coli* or *C. reinhardtii* platforms. At first glance it would seem that either the downstream box is less important than predicted, or that translation initiation is not limiting in the case of these genes. Further investigation of the pASap2 vector however, poses another possibility. In recent work by Dr Chloe Economou in the Purton lab the use of the pASap2 vector to create $atpA_{34}$*:GoI* chimeras has been shown to significantly decrease accumulation of the Cpl-1 lysin by increasing rates of proteolytic degradation (data not shown). The root of this

effect was concluded to be two fold. Firstly the $AtpA_{34}$:Cpl-1 chimera was shown to be rapidly degraded, likely due to the prevention of Cpl-1 correctly folding to its stable form. The full chimera, however, was only anticipated to be short lived as the stromal processing peptidase (SPP) was intended to remove the AtpA extension resulting in native, stable Cpl-1. This was indeed seen to occur, but only to a very low degree suggesting that access to the SPP recognition site was being blocked. A similar destabilisation effect was noted for the $AtpA_{34}$:Cpl-1 chimera in *E. coli* as shown in Figure 4.8.

In order to investigate possible causes for the instability seen, investigations were conducted into both the known structure of Cpl-1 and predicted structures for the pASap2 chimeric region. It was found that the N-terminus of Cpl-1 is located directly between the two domains of the protein, and is also in close proximity to a β-barrel central to the N-terminal catalytic domain (Figure 4.21). Predictions generated for the 57-residue pASap2 chimeric region indicate a compact globular structure comprising of four α-helices (Figure 4.22). It is conceivable that such a structure fused directly to the N-terminus of Cpl-1 could prevent correct folding, and therefore destabilise the protein sufficiently to trigger the degradation observed for the $AtpA_{34}$:Cpl-1 chimera. It is noted that the C-terminus is also located in this region; however, both the HA epitope tag and the flexi-linker region of the fusion lysins discussed below seem to be sufficiently unstructured to avoid destabilising Cpl-1.

To date, no crystal structures are available for Lys16 or Gp20 and thus comment on whether a similar destabilisation is taking place cannot be made. As such it is entirely possible that the incorporation of an endogenous downstream box did in fact increase translation initiation rates in the cases of *lys16* and *gp20,* only for the effect to be masked by an increased level of proteolytic degradation. The potential for further investigation into this area is discussed in the final chapter.

Figure 4.21 – The crystal structure of Cpl-1 shows the N-terminus to be buried between the two domains

The N-terminus of Cpl-1 is shown to be sandwiched between the two domains and is also closely associated with the β-barrel structure at the centre of the N-terminal catalytic domain. This protein structure was retrieved from the Protein Data Bank (PDB identifier 2G8J), and visualised using Jmol.

Figure 4.22 – A structural prediction of the pASap2 chimeric region shows a compact globular structure

The 57 residue chimeric region of pASap2 is predicted to form a compact globular structure formed of four α-helices. Structural predictions were generated by the Scratch Protein Predictor, and visualised using Jmol.

4.3.2 Implications of fusion protein development for improved expression and multi-specific enzyme production

A more general issue with the pASap2 construct may lie in the addition of a partial domain fragment to a stable protein. Instability could then be the result of either interference with folding of the target protein, or simply the presence of exposed hydrophobic residues in the N-terminal extension directly attracting proteases. It was therefore decided to move from an asymmetric chimera to a full bi-modular fusion protein. The gene for the highly expressed lysin Cpl-1 was fused upstream of the GoI with the additional goal of creating chimeric lysins effective against several bacterial pathogens. Three fusion genes were built, each comprising Cpl-1 followed by a 14-mer linker designed to form a random coil secondary structure, followed by either Lys16, Gp20, or a positive control, Pal.

4.3.2.1 *Effect on protein accumulation in* E. coli *and* C. reinhardtii

Despite the apparent high levels of expression observed for the *cpl-1:lys16* and *cpl-1:pal*, direct comparison with the corresponding singlet genes is not possible due to the use of the novel *psaA* promoter/ 5' UTR for these fusion constructs. However, the *cpl-1:gp20* construct was expressed using the original pASap1 vector under the control of the *atpA* promoter/ 5' UTR so allowing direct comparison with singlet *gp20* expression. When expressed in *E. coli*, *gp20* in isolation is undetectable by western blot analysis, but when *gp20* is expressed as part of a *cpl-1* fusion accumulation is greatly increased. It should be noted that although levels of the full-length fusion are low, the far more abundant C-terminal fragments all appear to contain full-length copies of the Gp20 protein, as based on their apparent mass. This suggests that Gp20 is sufficiently stable in *E. coli* to allow accumulation, and thus the previous absence of recombinant product is unlikely to be due to proteolysis. It is thus concluded that the creation of a bi-modular fusion lysin did improve expression of *gp20* in *E. coli.*

Any boost in expression observed in *E. coli* was not translated to expression in the *C. reinhardtii* chloroplast. The only fusion lysin construct showing any detectable expression was the *cpl-1:pal* positive control, and this was shown to be expressed at lower levels than singlet *cpl-1* under the same promoter/ 5' UTR.

4.3.2.2 ***Future prospects of fusion lysins for improved expression***

On the basis that *cpl-1* could be expressed to high levels as a single gene, and the fusion constructs faithfully reproduce the *cpl-1* translation initiation region, it is no longer thought that translation initiation is the limiting factor in the expression of *lys16* and *gp20*. If translation initiation is assumed to be occurring, the absence of product is likely due to either stalling of the translation elongation complex, or rapid turnover of the final product. The accumulation observed for Cpl-1:Pal shows that such fusions are not inherently unstable, and the predicted stabilities of Gp20 and Lys16 in the *C. reinhardtii* chloroplast suggest that stalled translation elongation is the most probable explanation for the lack of detectable fusion product.

However, as in the case of the chimeric $AtpA_{34}$ lysins, it is unclear how accessible the N-termini of Lys16 and Gp20 are, and thus how an N-terminal fusion would affect the folding and stability of the proteins. If the lack of detectable protein in *C. reinhardtii* is the result of proteolytic degradation, then a possible solution may be to redesign the linker region either by extension of the current sequence (as discussed above), or possibly a complete redesign to a more ridged GPGP hinge motif as used by Sun *et al* (Sun *et al.*, 2003). An alternative approach abandoning any bi-specific activity would be to add the stromal processing peptidase site as used in the pASap2 vector. Again assuming that N-terminal fusions do aid translation, this would allow for the fusion to be expressed as a single protein, then immediately be cleaved at the linker resulting in two stable lysins. Such an approach is discussed further in Chapter six.

If, however, rapid turnover is not the problem, then we can assume that translation is initiating and progressing through the *cpl-1* and linker portions of the transcript, but is then prematurely terminating during the course of the *lys16* and *gp20* sequences. As the HA tag is situated at the C-terminus this would result in no detectable accumulation on a western blot analysis, and give the impression that the gene was failing to be expressed altogether. Whether or not this is occurring could potentially be investigated in two ways. The first would be to modify the detection strategy to allow for C-terminally truncated (thus non-HA tagged) fusions to be identified. As the assumption is that translation stalling is

occurring after synthesis of the Cpl-1 moiety this could be achieved by purifying crude extract using the Cpl-1 DEAE ion exchange chromatography protocol and running the concentrated elutant on a protein gel. The presence of a specifically eluted protein would show progression of the elongation complex through the *cpl-1* portion of the transcript, and its mass would give a rough indication of stalling position. Similarly an N-terminal FLAG epitope tag or similar could be added to allow direct detection of C-terminal truncations by western blot. However, both of these processes rely on the partial fusion protein being sufficiently stable such that it can be detected. An opposing strategy would be to analyse translation elongation directly by conducting a polysome analysis. By this method direct data could be collected on ribosome position throughout the transcript, and thus conclusions could be drawn on firstly whether translational stalling was occurring, and if so, precisely what section of the transcript sequence was causing it. Due to time constraints these experiments could not be conducted. However, as a practical demonstration of the novel Codon Usage Optimizer software, work was moved directly to investigating whether improved codon optimisation could avoid such ribosome stalling events.

4.3.2.3 ***Future prospects for bi-functional lysins***

Activity assays were conducted for the three fusion lysins synthesised in *E. coli,* Cpl-1:Lys16, Cpl-1:Gp20, and Cpl-1:Pal, and the *C. reinhardtii* synthesised Cpl-1:Pal. Of these four only the *E. coli* synthesised Cpl-1:Pal displayed definitive functionality. This activity is, however, called into question: the inactive Cpl-1:Pal synthesised in the *C. reinhardtii* chloroplast is expressed as a single protein, whereas the active *E. coli* synthesised Cpl-1:Pal is shown to be severely fragmented. It is likely that the *E. coli* activity observed is due to a truncated form of the fusion protein as opposed to full length Cpl-1:Pal. It is thus concluded that this fusion lysin model does not produce active enzymes.

Given the known functionality of Cpl-1 and Lys16 in isolation, the linker region is likely to be the cause of this loss of activity. The flexi-linker was designed to form a random coil, allowing independent activity of the two enzymes without steric hindrance. It appears that this has only been partially successful: the linker is most likely to be forming a random coil, as evidenced by the stable accumulation seen

for *C. reinhardtii* synthesised Cpl-1:Pal. A structured peptide attached to the C-terminus of Cpl-1 would almost certainly disrupt folding of Cpl-1, which does not appear to be the case. Successful choline binding in purification also supports correct folding of individual modules; however, the lack of peptidoglycan hydrolase activity indicates that the two enzymes are sterically hindering one another. It is possible that the linker is not long enough to allow correct orientation of the lysins relative to their substrate. It is known, for example, that Cpl-1 dimerisation around the cell binding domain is required for full catalytic activity; this process could easily be blocked by the presence of the Pal N-terminal domain. An interesting avenue for future investigation into fusion lysins would thus be to investigate the effects of extending the linker region.

4.3.3 Redesign of *gp20* and issues raised

Based on the assumption that translation elongation was the limiting factor in expression of *gp20* and *lys16*, and in light of the bioinformatics investigations presented in Chapter five, *gp20* was redesigned following a novel codon usage strategy. However, despite this redesign no protein accumulation was seen in either *E. coli* or the *C. reinhardtii* chloroplast. Assuming that the lack of Gp20 accumulation is not due to degradation, the implication is thus that codon based ribosome stalling is not the limiting factor in *gp20* expression. Herein lies the issue with investigating such a multi-dimensional problem. Even if elongation could now be claimed to occur entirely unhindered by codon bias or other factors such as mRNA folding, there is still a multitude of other issues that could be blocking expression. A potential next step would thus be to take the redesigned *gp20* gene and repeat the experiments above to see if a combination approach of both improved translation initiation and elongation could be sufficient to allow detectable expression of *gp20* in the *C. reinhardtii* chloroplast.

4.3.4 The on-going potential for expression of *lys16* and *gp20* in *C. reinhardtii*

The initial aim of this project was to investigate the potential of the lysins Lys16 and Gp20 to be produced in the *C. reinhardtii* chloroplast. Despite various approaches, neither gene has shown any detectable protein accumulation using this platform. Though useful as a model for investigating failed recombinant gene

expression, the difficulty in achieving product accumulation for either gene suggests that these are perhaps not targets worth pursuing. In the case of *lys16* this is certainly the case; there are dozens of other potential *S. aureus* phage lysins, and the unique selling point of *C. reinhardtii* potentially producing a more soluble product than that seen in *E. coli* is somewhat superfluous if no expression is observed. The same cannot be said, however, for the *P. acnes* phage lysin Gp20. At the start of this project only a single *P. acnes* phage had been sequenced and annotated, making Gp20 unique. 15 such lysins have now been identified, but Gp20 is paradoxically still unique, as all 15 of these lysins show a remarkably high degree of sequence identity, thus rendering them virtually interchangeable. The target, *P. acnes,* also represents a unique opportunity in that therapeutics against it can be seen as duel-use technology, spanning both the pharmaceutical and cosmetic industries. The potential rewards of achieving expression of this gene are therefore worth continued investigation.

Chapter 5

A bioinformatics investigation into codon and codon pair use in the *C. reinhardtii* chloroplast

5.1 Introduction

5.1.1 Rationale for investigating codon usage

As was discussed in Chapter one, a considerable obstacle in the adoption of the *C. reinhardtii* chloroplast as a recombinant protein platform is that of low protein yield, encompassing both low expressing-, and apparently non-expressing genes. In the preceding chapter, translation initiation was optimised by the recreation of two translation initiation regions known to give high expression: the *atpA* 5' UTR leading into the first 104 bp of the *atpA* coding region, and the *atpA* 5' UTR leading into the full length *cpl-1* coding sequence. Though both are shown to be highly-expressing in their 'native' setting, neither of these leaders was able to bring expression of *gp20* or *lys16* to detectable levels of protein accumulation, despite positive controls in each case showing detectable product. The conclusions drawn from these data are that translation initiation is likely to be occurring in the cases of the *gp20* and *lys16* constructs, whereas translation elongation is stalling and thus detectable levels of product are not seen (although it must be acknowledged that the absence of detectable protein could also be a result of rapid turnover). The most influential factor in translation elongation is thought to be the codon usage of the transgene. This will be the focus of this chapter.

5.1.2 An introduction to codon usage

The codon usage of a gene simply refers to the choice in codons employed for each multi-coded amino acid residue due to the redundancy of the genetic cypher. Although all synonymous codons are capable of encoding the same amino acid, it has been shown that many organisms show a bias towards using certain codons (Sharp and Li, 1987; Welch *et al.*, 2009b). The codon optimisation of a transgene thus describes the codon usage of the recombinant gene in relation to the preferences, or bias, of the host organism. 'Favourable' codons are ones that appear often in highly expressing endogenous genes, and are believed to be translated more readily due to factors including relative abundance of tRNAs, recharge rate of different isoacceptors, and thermodynamic stability of the tRNA:ribosome complex. Conversely, 'unfavourable' codons are rarely found in native genes, and are thought to contribute to ribosome pausing or even complete stalling of the elongation complex.

The degree to which a transgene complies with its host's codon bias is referred to as the codon adaptation of the gene, and is expressed numerically in the form of a codon adaption index (CAI). This is calculated by taking the geometric mean (the *n*th root of the product of *n* numbers) of the relative weightings for all codons in a gene:

$$g\text{CAI} = \left(\prod_{i=1}^{n} w_i \right)^{1/n}$$

These weightings are typically collated in the form of a codon usage table, with codon weightings based on a collection of protein-coding genes from the host organism. These can be a highly expressing subset, or indeed the entire library of predicted protein-coding genes in a particular genome. The occurrence of each codon is then transformed to give a specific weight, w_i, where f_i is the codon frequency and f_j is the frequency of a synonymous codon:

$$w_i = \frac{f_i}{\max(f_j)}$$

As such, a specific codon weight can be described as the frequency ratio of the specified codon to the most frequent synonymous codon. A codon usage table supplying codon weights allows for a gene to be designed following the same codon preferences as seen in the host organism, theoretically allowing for improved translation elongation relative to a non-optimised gene. Such optimisation is typically conducted *in silico* by an optimisation program, with such services frequently being offered commercially in conjunction with gene synthesis.

5.1.3 Issues with the current *C. reinhardtii* chloroplast codon usage table

Codon optimisation is routinely employed when designing genes for expression in the *C. reinhardtii* chloroplast, showing varying degrees of success as discussed in Chapter one (1.4.5). The most widely used codon usage table employed by the *C. reinhardtii* chloroplast community was derived by the Kazusa Institute as part of a project to generate codon usage data for a wide range of genomes (Table 5.1) (Nakamura *et al.*, 2000).

Table 5.1 – The Kazusa Institute *C. reinhardtii* chloroplast codon usage table

Data is displayed in terms of absolute frequency, and also as frequency per thousand codons. The important point to note in this case is not the codon data itself, but the number of coding sequences from which they have been generated.

***chloroplast Chlamydomonas reinhardtii* [gbpln]: 93 CDS's (26731 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

UUU	33.4(894)	UCU	17.0(455)	UAU	24.6(657)	UGU	7.6(203)
UUC	17.1(456)	UCC	2.8(74)	UAC	10.0(266)	UGC	1.5(39)
UUA	77.7(2078)	UCA	22.0(588)	UAA	2.9(78)	UGA	0.1(3)
UUG	4.3(114)	UCG	4.0(107)	UAG	0.4(12)	UGG	13.5(361)
CUU	14.3(383)	CCU	15.5(414)	CAU	10.1(270)	CGU	32.4(866)
CUC	1.0(28)	CCC	3.4(90)	CAC	8.8(235)	CGC	4.1(110)
CUA	6.4(170)	CCA	23.6(630)	CAA	38.4(1026)	CGA	3.4(90)
CUG	3.7(99)	CCG	2.4(63)	CAG	4.1(110)	CGG	0.5(14)
AUU	51.4(1374)	ACU	24.4(651)	AAU	42.1(1126)	AGU	16.0(428)
AUC	8.2(219)	ACC	5.1(135)	AAC	17.7(472)	AGC	5.4(144)
AUA	6.9(184)	ACA	32.4(865)	AAA	69.1(1847)	AGA	5.3(143)
AUG	22.3(596)	ACG	3.9(103)	AAG	6.2(167)	AGG	0.9(23)
GUU	29.3(783)	GCU	34.0(908)	GAU	25.3(676)	GGU	44.0(1177)
GUC	2.5(68)	GCC	5.9(159)	GAC	9.8(263)	GGC	6.4(172)
GUA	26.0(696)	GCA	20.7(554)	GAA	41.1(1098)	GGA	8.6(229)
GUG	5.6(149)	GCG	3.3(88)	GAG	5.7(152)	GGG	3.7(99)

Coding GC 33.72% 1st letter GC 44.40% 2nd letter GC 37.35% 3rd letter GC 19.40%

The issue with the Kazusa codon usage table is the quoted number of coding sequences (CDS) used to generate the table, a figure of 93. There are only 68 unique protein-coding genes to be found in the *C. reinhardtii* chloroplast genome[14], and as such the Kazusa table has clearly incorporated an extra 25 sequences of unknown origin, possibly from duplicate versions of the same genes, or from rRNA or tRNA genes. Without knowing the precise sequences used to form the table it is not possible to comment on the degree by which the table may have been skewed by these extra CDS; however, it is clear that such a table will be non-representative. For reference, the *C. reinhardtii* chloroplast genome is displayed in Figure 5.1.

[14] http://www.ncbi.nlm.nih.gov/nuccore/41179002

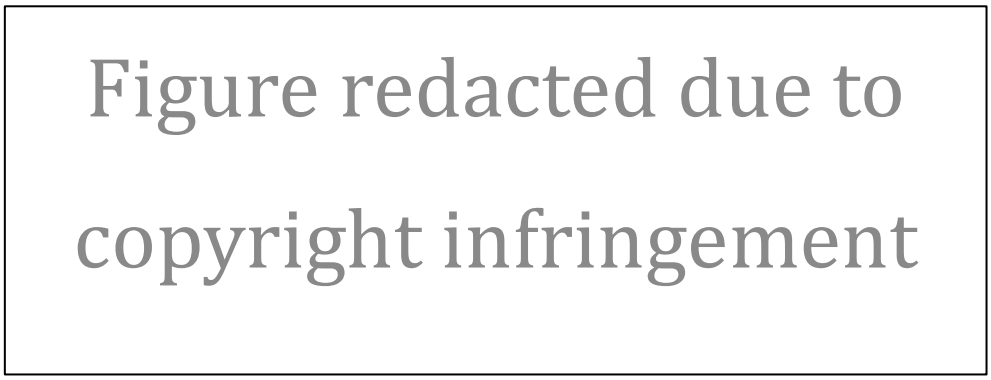


Figure 5.1 – The *C. reinhardtii* chloroplast genome

The inner circle shows *Bam*HI and *Eco*RI restriction fragments, the second concentric circle indicates seven overlapping BAC clones that span the genome, and the third circle shows genes and ORFs of unknown function, as of 2002. The outer circle shows genes of known or presumed function, with sequenced or hypothesized introns shown in olive green. Genes are color coded by function, as shown at bottom. Figure and legend adapted from (Maul *et al.*, 2002).

A further issue with conventional codon optimisation using the above table is that it does not take codon pairing into consideration. It has been shown that in addition to an individual codon preference, organisms also show partiality for particular codon pairs. The reasoning behind this notion is based on the interactions that occur between pairs of tRNAs while occupying the A and P sites of the ribosome (Smith and Yarus, 1989). Studies in *E. coli* have demonstrated 'favourable' codon pairs to be translated more rapidly than 'unfavourable' pairs, in the same manner as for individual codon preferences (Boycheva *et al.*, 2003; Coleman *et al.*, 2008). To date, no research has been conducted on whether the chloroplast ribosome (*C. reinhardtii* or otherwise) also displays codon pair bias.

5.1.4 The Codon Usage Optimizer as a novel tool for codon analysis and optimisation

The Codon Usage Optimizer (CUO) is a novel piece of software developed in the Purton lab by Khai Kong Jien, an undergraduate project student. Its primary function is as a gene optimiser, but it also has the ability to analyse codon- and codon pair frequencies from a defined library of genes. The latter function can be used either to construct bespoke codon usage tables or as a basic research tool to investigate the nature of such biases. During the course of this chapter, this feature is used heavily to analyse both the *C. reinhardtii* chloroplast native library of 68 distinct protein-coding genes, and also the growing number of recombinant genes synthesised by the Purton lab.

5.1.5 General strategy for investigation

It is now generally acknowledged that transgenes intended for expression in the *C. reinhardtii* chloroplast should be codon-optimised in order to give maximum recombinant yield (Purton *et al.*, 2013; Rosales-Mendoza *et al.*, 2012); however, little analytical work has actually been conducted into codon usage in *C. reinhardtii*, and apparently none at all on codon pair usage. This chapter is intended as a preliminary investigation into this field.

Following on from the investigations of the previous chapter, the focus of this work will be on the reliability of expression as opposed to protein yield. The advantages of such a strategy are twofold: firstly, it allows for a greatly simplified approach suitable for a preliminary investigation, as expression can be defined as a discreet on/off variable (as opposed to a sliding scale), and secondly, as discussed in Chapter one, the issue of expression reliability can be seen as more pressing at this point, in terms of developing the *C. reinhardtii* chloroplast as a recombinant protein platform. As such, the analyses presented below are less orientated towards ideal codon- and codon pair usage to maximise expression, and more towards preventing unfavourable usage, which may lead to the blocking expression altogether. To this end, the reference set of genes used throughout is the complete library of 68 protein-coding genes, as opposed to the high expressing subset used previously for the redesign of *gp20* (see 4.2.5).

In order to investigate how codon usage may affect transgene expression in such an absolute manner, the data is generally analysed in a two-stage approach. The first of these stages is the direct analysis of native genes expressed in the *C. reinhardtii* chloroplast, with the aim of identifying incidences where particular codons or codon pairs are specifically avoided. In the second stage, this information is then related back to the recombinant genes thus far introduced into the *C. reinhardtii* chloroplast genome in the Purton lab in order to ascertain whether features avoided in the native context give rise to non-expressing genes.

5.1.6 Specific aims and objectives

1. To investigate the codon- and codon pair distribution in the complete set of 68 protein-encoding *C. reinhardtii* chloroplast genes in order to deduce if global biases are indeed present.

2. To apply global codon- and codon pair scoring generated by the CUO to recombinant transgenes expressed in the Purton lab in order to investigate if there is a correlation between global codon pair use and absolute expression.

3. To investigate the phenomenon of Zero Scoring Codon Pairs (ZSCPs: codon pairs that are completely unseen in the *C. reinhardtii* chloroplast) in relation to corresponding predicted values based on individual codon and amino acid usage models.

4. To apply the concepts of ZSCPs to recombinant genes in order to analyse possible lethal pair combinations.

5.2 Results

Due to the bioinformatic nature of the data presented, and the organic manner by which these investigations have developed, the results will be presented as a series of ten hypotheses to be investigated as the chapter progresses. For reference, they are listed below:

Hypothesis One. The protein-coding genes of the *C. reinhardtii* chloroplast show a defined codon bias.

Hypothesis Two. The protein-coding genes of the *C. reinhardtii* chloroplast show a defined codon pair bias.

Hypothesis Three. Global codon usage influences recombinant gene expression.

Hypothesis Four. Global codon pair usage influences recombinant gene expression.

Hypothesis Five. Local codon pair usage influences recombinant gene expression.

Hypothesis Six. All Zero Scoring Codon Pairs (ZSCPs) observed for the *C. reinhardtii* chloroplast are explicitly avoided.

Hypothesis Seven. The failure of non-expressing genes can be explained by the presence of unexpectedly unseen ZSCPs.

Hypothesis Eight. Regions of low codon pair usage are responsible for non-expressing recombinant genes.

Hypothesis Nine. ZSCPs are conserved across a panel of related green algal chloroplast genomes.

Hypothesis Ten. The available tRNA pool reflects the relative codon preferences seen in the *C. reinhardtii* chloroplast genome.

5.2.1 Hypothesis One – The protein encoding genes of the *C. reinhardtii* chloroplast show a defined codon bias

It is clear from previous work (Heitzer *et al.*, 2007; Nakamura *et al.*, 2000) that the *C. reinhardtii* chloroplast does show codon bias; however, steps were taken to confirm that the data generated by the CUO was in agreement with this position.

The CUO program was used to analyse the 68 protein-coding genes in the *C. reinhardtii* chloroplast. (There are in fact 69 such genes, but psbA is duplicated on the inverted repeat regions so will only be included once in these analyses). The observed dataset created will henceforth be known as *o(dataset1*). A tabulated breakdown of the datasets used in this analysis can be found in Appendix u. Data from such analyses using the CUO is produced in the form of a codon summary table, a codon weight table and a far more extensive codon pair table encompassing the 3904 possible combinations. Each of these can be converted into a Microsoft Excel format suitable for further analysis. The codon summary and codon pair tables include data for each codon/ pair relating to total occurrence and occurrence per thousand codons/ pairs. The codon weight table contains weightings as described above. These are in the form of relative weightings of each codon in relation to the amino acid encoded, and although weighting is important for the optimisation of sequences, it is less helpful for usage analysis so will not be analysed further following this hypothesis.

It is clear from a visual analysis of the weighted codon usage table shown in Figure 5.2 that a strong codon bias is present in the *C. reinhardtii* chloroplast. However, to illustrate this point in an objective manner, a null hypothesis was suggested: codon use in the *C. reinhardtii* chloroplast is entirely dependent on amino acid ($\alpha\alpha$) usage. To test this hypothesis, the frequencies of each codon were taken from *o(dataset1)* and compared to a predicted set of codon frequencies taking only amino acid usage into account. This predicted dataset was named *p(dataset1)*.

p(dataset1) was produced by first calculating absolute frequencies for each amino acid by dividing the total occurrence of that $\alpha\alpha$ by the total number of codons in the dataset (25,392), minus the 68 stop codons (25,324).

$$p(\alpha\alpha_1) = o(\alpha\alpha_1)/25324$$

This figure was then split between each synonymous codon, *n*, for the amino acid in question to give predicted absolute frequencies for each codon assuming completely equal codon weighting.

$$p(C_1)_{abs.} = p(\alpha\alpha_1)/n$$

These absolute frequencies were then multiplied by the total number of non-stop codons to give expected codon occurrences assuming only $\alpha\alpha$ bias.

$$e(C_1) = p(C_1)_{abs.} * 25324$$

The codon occurrences from o(*dataset1)* and *p(dataset1)* were then plotted to allow visual analysis. Chart 5.1 shows the expected distribution of codon usage in a completely non codon-biased *C. reinhardtii* chloroplast genome, and the actual distribution seen. The radically different distribution observed strongly implies that the codon usage in the *C. reinhardtii* chloroplast is indeed biased.

A Pearson's χ^2 test for goodness-of-fit was conducted to give a statistical significance to the codon bias observed. The test returned a χ^2 value of 22100 (3 significant figures) with a p value of 0 (as rounded by Microsoft Excel) at 60 degrees of freedom. This confirms, to a high statistical significance, that even distribution of synonymous codons does not occur in the *C. reinhardtii* chloroplast, proving the presence of a codon bias.

a

Summary table:

Sequence length: 76176 Number of protein genes: 68 Total codons: 25392

Amino acid	Codon	Number	Frequency per thousand												
Phe	UUU	867	34.1	Ser	UCU	477	18.8	Tyr	UAU	668	26.3	Cys	UGU	175	6.9
	UUC	426	16.8		UCC	57	2.2		UAC	234	9.2		UGC	20	0.8
Leu	UUA	2034	80.1		UCA	594	23.4	STOP	UAA	64	2.5	STOP	UGA	0	0.0
	UUG	99	3.9		UCG	98	3.9		UAG	4	0.2	Trp	UGG	326	12.8
Leu	CUU	377	14.8	Pro	CCU	380	15.0	His	CAU	271	10.7	Arg	CGU	862	33.9
	CUC	14	0.6		CCC	43	1.7		CAC	222	8.7		CGC	66	2.6
	CUA	176	6.9		CCA	576	22.7	Gln	CAA	1016	40.0		CGA	91	3.6
	CUG	47	1.9		CCG	53	2.1		CAG	81	3.2		CGG	6	0.2
ile	AUU	1340	52.8	Thr	ACU	651	25.6	Asn	AAU	1185	46.7	Ser	AGU	424	16.7
	AUC	204	8.0		ACC	86	3.4		AAC	468	18.4		AGC	117	4.6
	AUA	182	7.2		ACA	859	33.8	Lys	AAA	1940	76.4	Arg	AGA	141	5.6
Met	AUG	540	21.3		ACG	95	3.7		AAG	122	4.8		AGG	19	0.7
Val	GUU	699	27.5	Ala	GCU	873	34.4	Asp	GAU	646	25.4	Gly	GGU	1130	44.5
	GUC	19	0.7		GCC	100	3.9		GAC	202	8.0		GGC	115	4.5
	GUA	675	26.6		GCA	520	20.5	Glu	GAA	1062	41.8		GGA	202	8.0
	GUG	96	3.8		GCG	78	3.1		GAG	87	3.4		GGG	91	3.6

0 unidentified codons.

Nucleotide frequency: A=26777 U=24963 G=12222 C=12214 X=0

GC Content: 32.08% 1st letter GC: 42.83% 2nd letter GC: 36.72% 3rd letter GC: 16.68%

b

Amino acid	Codon	Weight									
Phe	UUU	1	Ser	UCU	0.803	Tyr	UAU	1	Cys	UGU	1
	UUC	0.491		UCC	0.096		UAC	0.35		UGC	0.114
Leu	UUA	1		UCA	1	STOP	UAA	---	STOP	UGA	---
	UUG	0.049		UCG	0.165		UAG	---	Trp	UGG	---
Leu	CUU	0.185	Pro	CCU	0.66	His	CAU	1	Arg	CGU	1
	CUC	0.007		CCC	0.075		CAC	0.819		CGC	0.077
	CUA	0.087		CCA	1	Gln	CAA	1		CGA	0.106
	CUG	0.023		CCG	0.092		CAG	0.08		CGG	0.007
ile	AUU	1	Thr	ACU	0.758	Asn	AAU	1	Ser	AGU	0.714
	AUC	0.152		ACC	0.1		AAC	0.395		AGC	0.197
	AUA	0.136		ACA	1	Lys	AAA	1	Arg	AGA	0.164
Met	AUG	---		ACG	0.111		AAG	0.063		AGG	0.022
Val	GUU	1	Ala	GCU	1	Asp	GAU	1	Gly	GGU	1
	GUC	0.027		GCC	0.115		GAC	0.313		GGC	0.102
	GUA	0.966		GCA	0.596	Glu	GAA	1		GGA	0.179
	GUG	0.137		GCG	0.089		GAG	0.082		GGG	0.081

Figure 5.2 – Codon usage (a) and weight tables (b) for the 68 unique protein-coding genes of the *C. reinhardtii* chloroplast show a definite codon bias

A weighting of 1 represents the most abundant codon for the amino acid in question, with all other weightings in the group expressing the proportional share of usage relative to the most abundant codon. Direct visual analysis of either table suggests an uneven distribution of synonymous codons for each αα, implying a well-defined codon bias.

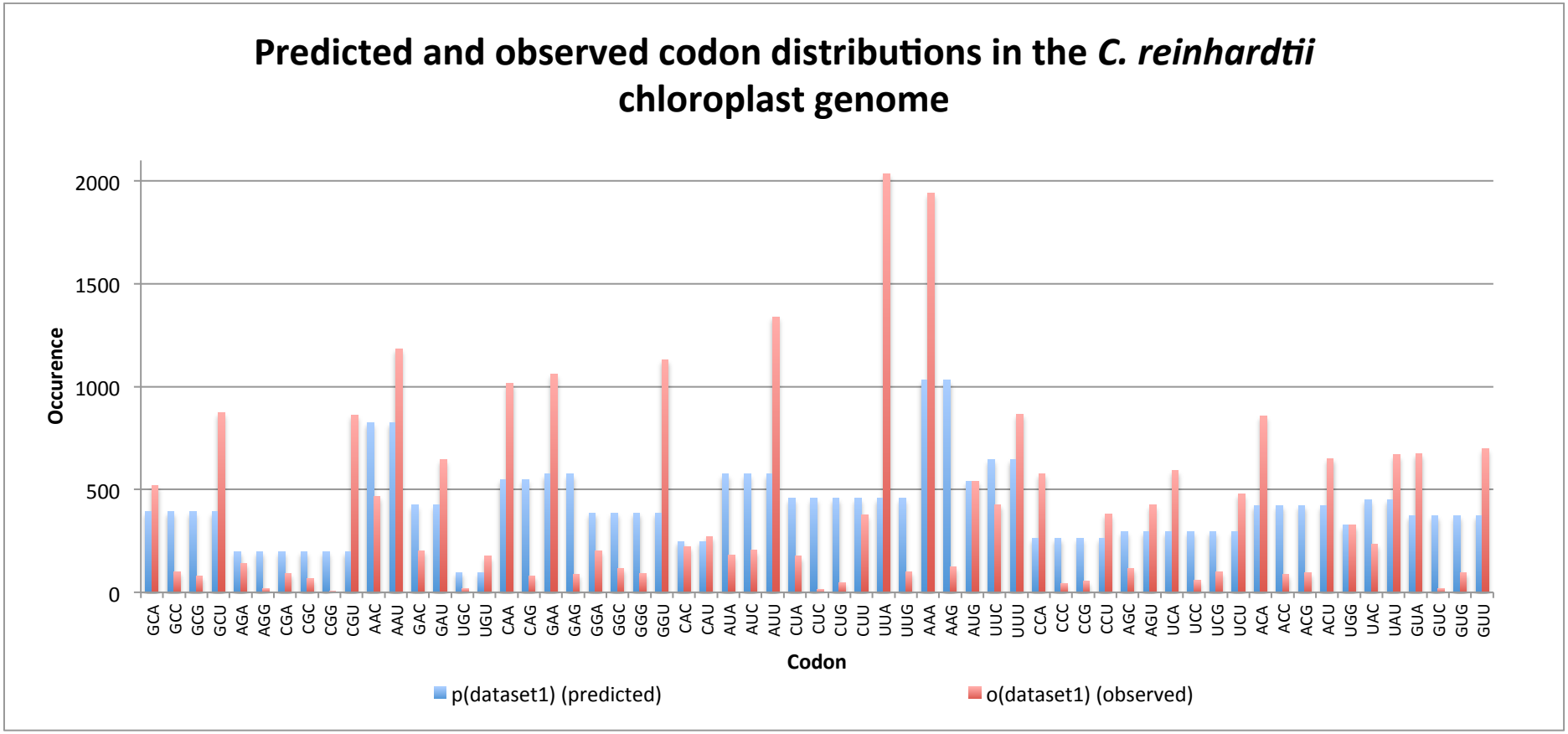


Chart 5.1 – Predicted and observed codon distributions in the *C. reinhardtii* chloroplast genome show marked deviations

Predicted codon occurrences from *p(dataset1)* are generated assuming that there is no codon bias in the *C. reinhardtii* chloroplast genome and that codon use is spread evenly amongst all synonymous codons. By comparison with the observed codon distribution shown in *o(dataset1)*, it is clear that this assumption is incorrect and there is strong preference for certain codons over others, demonstrating a codon bias.

5.2.2 Hypothesis Two – The protein-coding genes of the *C. reinhardtii* chloroplast show a defined codon pair bias

Prior to this investigation, there was no published work that explored whether a bias is shown towards codon pairing in the *C. reinhardtii* chloroplast genome, or indeed the plastid genome of any other species. In this section, *o(dataset1)* is analysed in the context of codon pairing, with the goal of discerning whether a bias exists that is not explained by either the combination effects of individual codon use, amino acid usage, or amino acid pairing.

The CUO is capable of generating graphical codon pair tables; however, three major factors prevent such tables from being appropriate for direct visual analysis in the same manner as for the individual codon weight tables above (5.2.1). Firstly, the data is presented independently of individual codon usage, making it impossible to separate genuine codon pair bias from the compound effect of individual codon preferences. Secondly, the data does not take into account relative frequency of amino acid or amino acid pair usage, and thirdly, given the 4096 pair combinations, the data are simply too large to efficiently analyse by eye. Instead, a similar approach was taken to that used in Hypothesis 1; a number of null hypotheses were proposed, and predicted datasets were generated for each hypothesis for comparison against *o(dataset1)*.

The first null hypothesis to be assessed was that codon pairing is purely a function of the absolute frequencies of the individual codons that make up that pair. A predicted dataset was generated, *p(dataset2)*, incorporating individual codon frequencies and hence, also total amino acid usage. Codon absolute frequencies were calculated in a similar manner to the $\alpha\alpha$ frequencies seen in Hypothesis 1, by dividing the total occurrence for each codon by the total number of codons in the dataset minus the 68 stop codons (25324). The resulting decimal generated for each codon reflects the probability of a particular codon being incorporated at any one position.

$$p(C_1)_{abs.} = o(C_1) \,/\, 25324$$

As this model assumes no codon pair preference, the incorporation of two sequential codons can be considered to be independent events. As such the probability of any pair can be considered to be the product of the individual codons' probabilities:

$$p(C_1 : C_2)_{abs.} = p(C_1)_{abs.} * p(C_2)_{abs.}$$

From the pair probabilities generated, an expected value of occurrence, $e(C_1{:}C_2)$, can then be calculated by multiplying the pair probability by the total number of codon pairs in the dataset. Each gene containing *n* codons can be considered to contain *n-2* codon pairs giving a total number of codon pairs as 25256 as illustrated in Figure 5.3.

$$e(C_1 : C_2) = p(C_1 : C_2)_{abs.} * 25256$$

The predicted *p(dataset2)* was plotted alongside the observed codon pairs seen in *o(dataset1)* (Chart 5.2), with the data ordered by ratio of observed to predicted values. As such, codon pairs seen less often than expected are clustered to the left, while those that are seen more often than predicted are clustered towards the right. It can be seen from the graph that *p(dataset2)* and *o(dataset1)* form distinct populations representing a considerable variance between the two datasets. This suggests that additional factors are acting on codon pair distribution relative to those considered by *p(dataset2).*

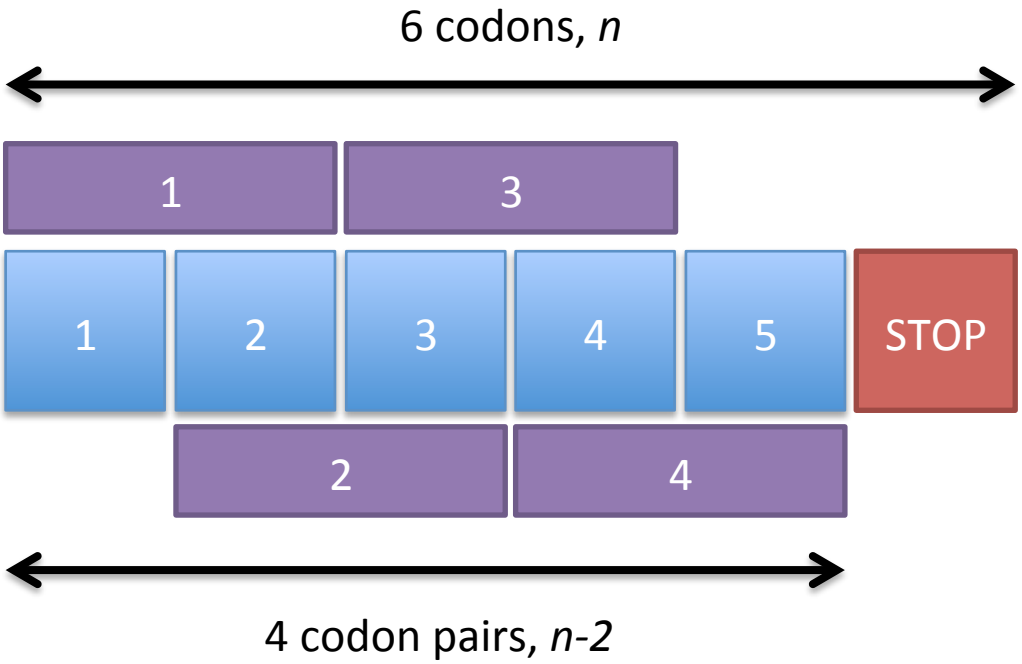


Figure 5.3 – A graphical illustration of the relationship between total codons and total codon pairs

For any gene the number of coding codon pairs will equal *n-2* where *n* is the total number of codons present. The total number of protein-coding codons can be seen to be *n-1*.

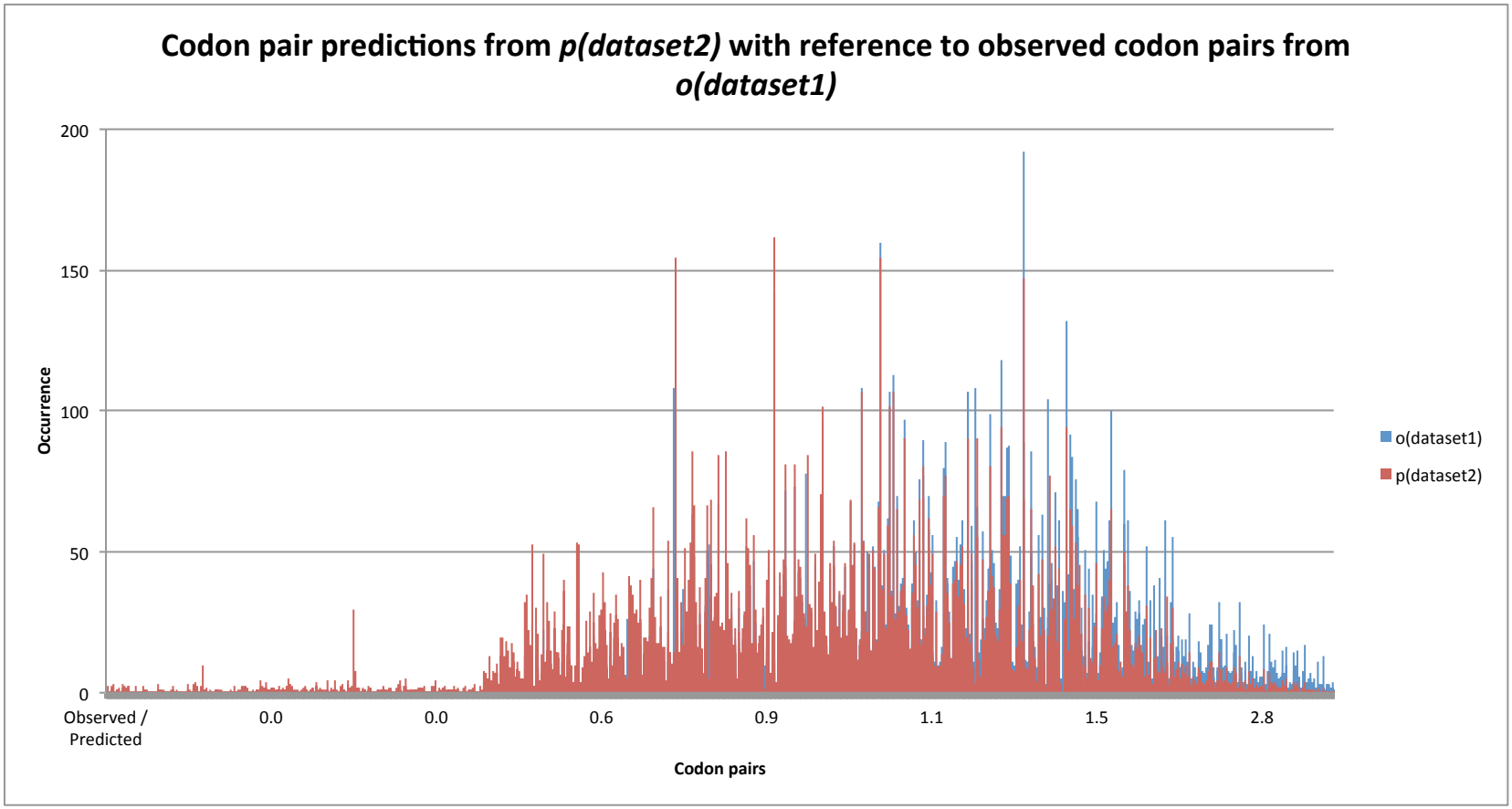


Chart 5.2 – Codon pair distribution for predicted and observed datasets *p(dataset2)* and *o(dataset1)*

Predicted and observed codon pair frequencies are plotted against codon pairs ordered by the ratio of observed to predicted occurrences. Although difficult to quantitatively assess, it is clear that there is considerable deviation between the predicted and observed datasets, suggesting that there are other factors influencing the selection of codon pairs. Groupings of over- and under- represented codon pairs are clearly evident.

The values for observed and predicted codon pair occurrences are useful for showing general trends in the data in regard to whether pairs are over- or under-represented; however, they are not optimised for analysis of the degree to which this is occurring as it relies on absolute values. To analyse how observed codon pair usage diverges from expected values in the absence of a codon pair preference, a direct comparison of *p(dataset2)* and *o(dataset1)*, Δ (codon pair use) was created:

$$\Delta(C_1 : C_2) = o(C_1 : C_2) \,/\, e(C_1 : C_2)$$

These data were then plotted on a $\log_2$ scale to provide a visual illustration of deviation from the expected values in the absence of a codon pair bias (Chart 5.3). For clarity, codon pairs are ordered by variance. Positive values represent the -fold increase in the observed frequency relative to predicted codon pair occurrence, *id est* are under-represented by the model. Negative values show the -fold decrease in observed frequency relative to the expected value, corresponding to over-representation. The use of a log scale allows both positive and negative variation to be displayed on the same scale.

It is clear from these data that the predicted dataset *p(dataset2)* shows considerable deviation from the actual codon pair use in the observed dataset *o(dataset1)*. The data also appear to be appreciably skewed to the right, indicating an excess of codon pairs that are under-represented by the model. It must be taken into account, however, that of the 3721 non-stopping codon pairs, 1142 are not seen in the chloroplast genome. For such cases Δ (codon pair use) will give a value of zero, and thus cannot be plotted on a log scale graph. As *p(dataset2)* does not contain any zero values, all 1142 zero data points would have appeared on the left hand side of the plot, balancing the skew to some degree.

A brief analysis of the data reveals that of the 2579 codon pairs observed in the *C. reinhardtii* chloroplast genome, 1910 (74.1 %) have a variance that falls between -0.5 and +2; that is to say, they are within one-fold variance relative to the expected value given by *p(dataset2)*. Of those remaining, 426 (16.5 %) are seen more than twice as often as expected, and 243 (9.40 %) are seen half as often.

Statistical significance is confirmed by the Pearson's χ^2 test for goodness-of-fit, with a χ^2 value of 5980 and a p value of 1.31×10^{-109} (3 s.f.) at 3720 degrees of freedom. The null hypothesis that that codon pairing is purely a function of the absolute frequencies of the individual codons that make up that pair has thus been disproved.

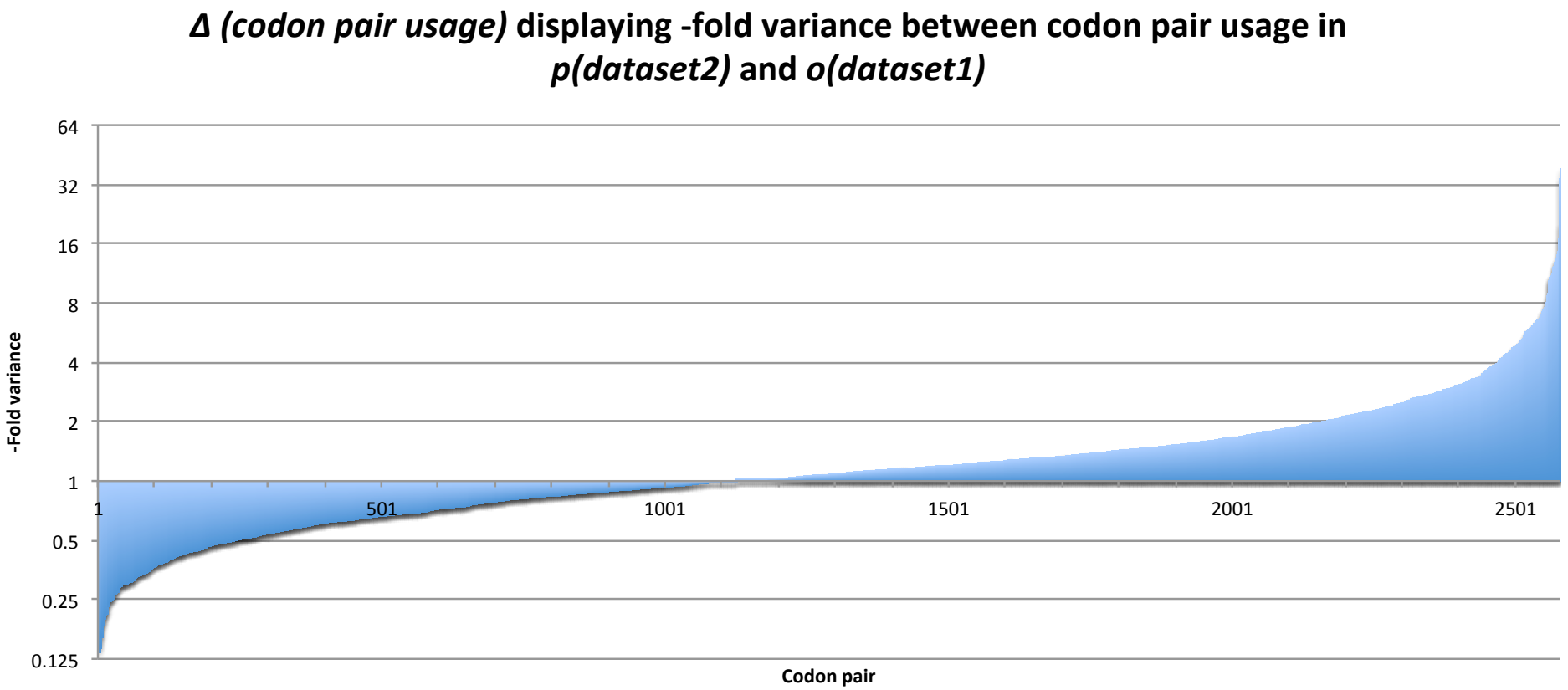


Chart 5.3 – A plot of Δ (codon pair usage) shows considerable -fold variance between *p(dataset2)* and *o(dataset1)* codon pair usage

These data show a considerable degree of variation between predicted and observed codon pair frequencies. Data appears to be skewed to the right (under prediction of codon pairs). This is partly due to the effects of zero scoring codon pairs in *o(dataset1)* which cannot be plotted on a log scale graph.

It is noted, however, that although *p(dataset2)* takes codon use and amino acid make-up into account, it shows no regard for αα pairings. Such pairings are important for protein structural motifs, so a further null hypothesis was proposed: amino acid pairing has no bearing on gene design in the *C. reinhardtii* chloroplast.

To assess this hypothesis, *p(dataset2)* was translated into an amino acid pair dataset *p(dataset2α)*, and this then compared to the actual amino acid pair occurrence seen by creating a Δ(αα pair use):

$$\Delta(\alpha\alpha_1) = o(\alpha\alpha_1) \,/\, e(\alpha\alpha_1)$$

These data are displayed in Chart 5.4. The variance seen is considerably lower than that seen in the codon pairing variance analysis of Chart 5.3, with only 1.5 % of αα pairs being more than 1-fold variance from the observed data. However, given the origin of the data, and also the fact that back-translating removes all codon bias, a far closer relationship would be expected if αα pairing was not influential, as confirmed by a significant χ^2 result of 853 and a p value of 3.95×10^{-35} (3 s.f.) at 399 degrees of freedom.

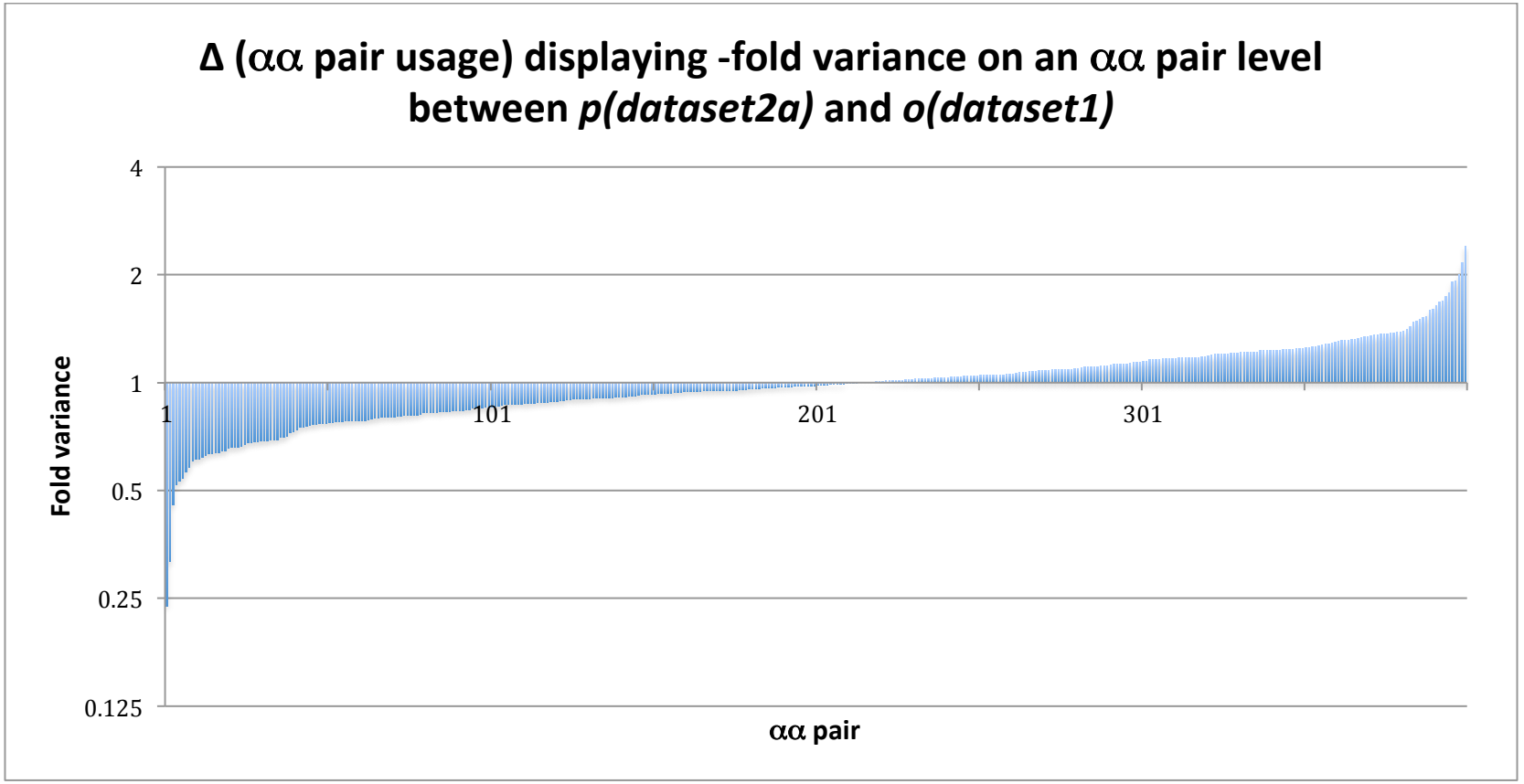


Chart 5.4 – Δ (αα pair usage) displaying variance on an αα pair level between *p(dataset2a)* and *o(dataset1)*

The data used to build *p(dataset2)* came directly from *o(dataset1)*; therefore if the model employed was capable of taking αα pairing into account, there should be no variance between the observed and predicted αα pair frequencies. This is clearly not the case, although the variance is much smaller than that seen for Δ (αα pair usage), suggesting that while αα pairing was contributing to the deviation seen in chart 5.3, it was not the only factor.

It is clear that *p(dataset2)* is flawed in that it does not take αα pairing into account, and thus variance between it and the observed values of *o(dataset1)* cannot be used as confirmation of codon pair bias. To refine the predicted model, a new dataset, *p(dataset3*) was generated using amino acid pairing frequencies as a base dataset and relative codon usage for each codon as a probability factor. By using amino acid pair data, both amino acid use and pairing is controlled, and combined with individual relative codon usage data, any significant deviation of the predicted dataset from the observed codon pairing values can only be explained by specific codon pair preference. (For the purpose of this argument, any other mRNA sequence motifs are considered to be acting, at least on a basic level, through codon pairing.)

The predicted codon pair outcome for *p(dataset3)* could not be generated from the absolute frequency as used in *p(dataset2)* as this probability factor already incorporates amino acid usage, and thus was inappropriate for an analysis already based on amino acid pair data. Instead a new probability factor was generated – p(Relative Codon Usage):

$$p(C_1)_{rel.} = o(C_1) \,/\, \sum o(C_{\alpha\alpha 1})$$

This gives a probability factor equivalent to the likelihood of a specific codon being used at any time for its specific amino acid. *p(dataset3)* was then generated using the following expression to give codon pair predictions weighted by the observed αα pair usage:

$$e(C_1 : C_2)_w = p(C_1)_{rel.} * p(C_2)_{rel.} * o(\alpha\alpha_1 : \alpha\alpha_2)$$

The dataset was verified by back-translating the predicted codon pairs into αα pairs and relating the predicted frequencies to the observed αα pairs seen in *o(dataset1)*. The two datasets were in complete agreement, indicating that any variance between codon pairs could only be due to a codon pair bias. To ensure individual codon bias had been correctly incorporated by the model, codon usage was back-calculated from codon pair data for the observed dataset *o(dataset1)* and the predicted *p(dataset3)*. With the exception of rounding errors totalling 0.129 %,

the codon usage data was in total agreement between the two datasets. With αα and individual codon bias corrected for, any variance must therefore be due to a codon pair bias.

Chart 5.5 shows predicted, *p(dataset3),* and observed, *o(dataset1),* codon pair occurrence as displayed for *p(dataset2)* in Chart 5.2. Relative to *p(dataset2)*, *p(dataset3)* is seen to show closer relation to the observed codon pair distribution; however, there is still considerable variation from the observed codon pair data.

As for *p(dataset2),* a variance analysis was produced for *p(dataset3)* and this is displayed against that of *p(dataset2)*. Only a very minor deviation between the two datasets is observed. This suggests that although the removal of αα pair bias from *p(dataset3)* has resulted in a model more closely resembling the observed values, a much larger source of bias affecting codon pair use remains. This is shown to be significant by a χ^2 result of 5030 and a p value of 3.74×10^{-43} (3 s.f.) at 3720 degrees of freedom.

Of the 2579 codon pair combinations seen in the *C. reinhardtii* chloroplast genome (after removal of zero scoring codons pairs), 1948 (75.5 %) have a variance that falls between -0.5 and +2, 423 (16.4 %) are seen more than twice as often as expected, and 208 (8.07 %) are seen half as often as expected. Again, it must be taken into account that the zero scoring codon pairs will skew the picture towards the over representation of codon pairs. The issue of zero scoring pairs does not lend itself to this form of analysis, but will be investigated further below.

The results here discussed have shown that there is indeed a highly statistically significant bias affecting both codon usage and codon pairing, neither of which can be attributed to amino acid usage, amino acid pairings (as could perhaps be dictated by structural motifs or steric hindrance issues), or the combined bias effect of the codon pair's component codons. The specific nature of this bias, in particular in regard to zero scoring codons, remains uncertain and is explored further in the following hypotheses.

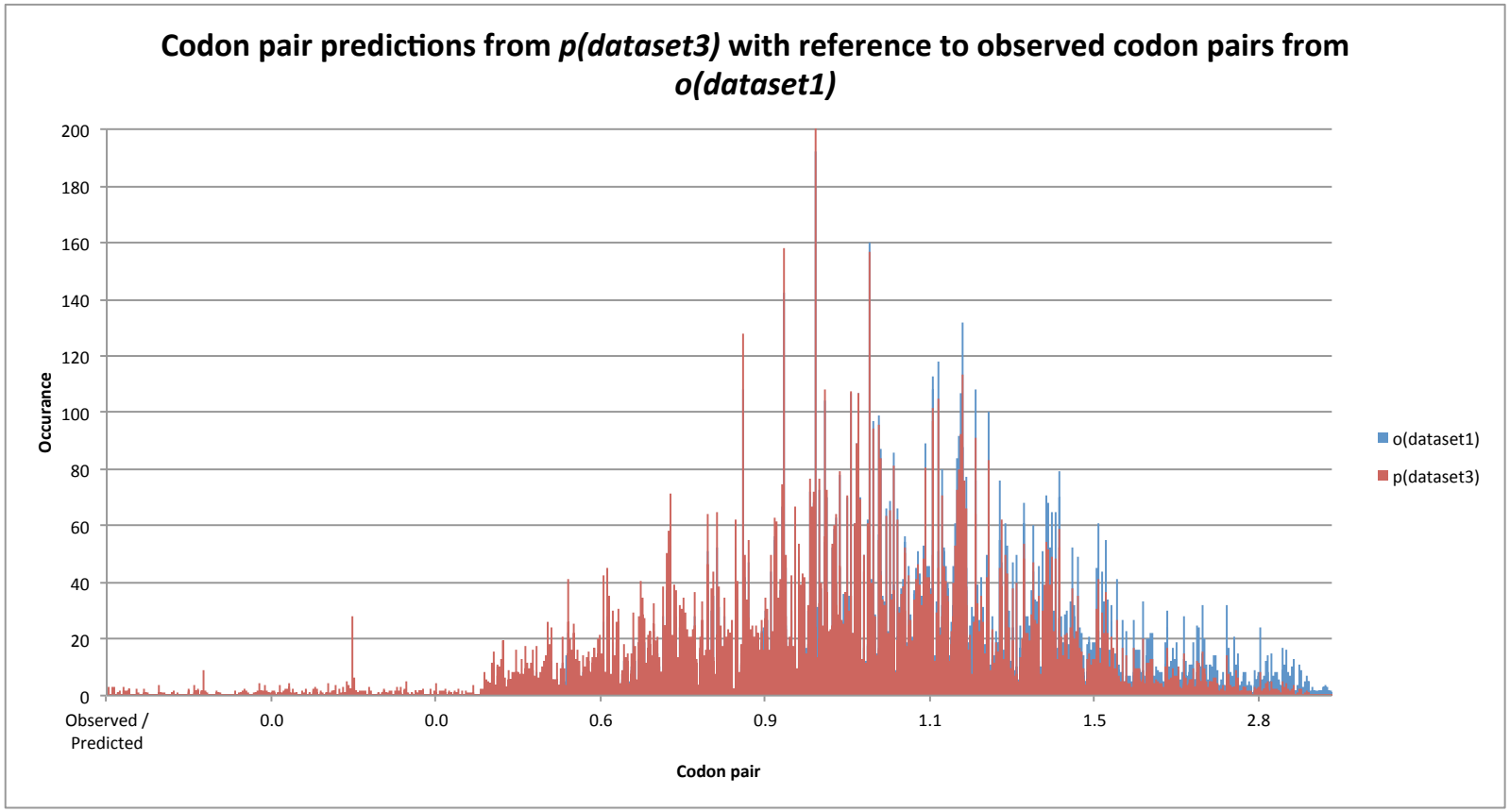


Chart 5.5 – Codon pair distribution for the predicted dataset *p(dataset3)* relative to *o(dataset1)* displays substantial variation

Predicted and observed codon pair frequencies are plotted against codon pairs ordered by the ratio of observed to predicted occurrences. Relative to *p(dataset2)* (Chart 5.2), predicted and observed codon pair occurrences are more closely aligned, but there is still a notable variance between them.

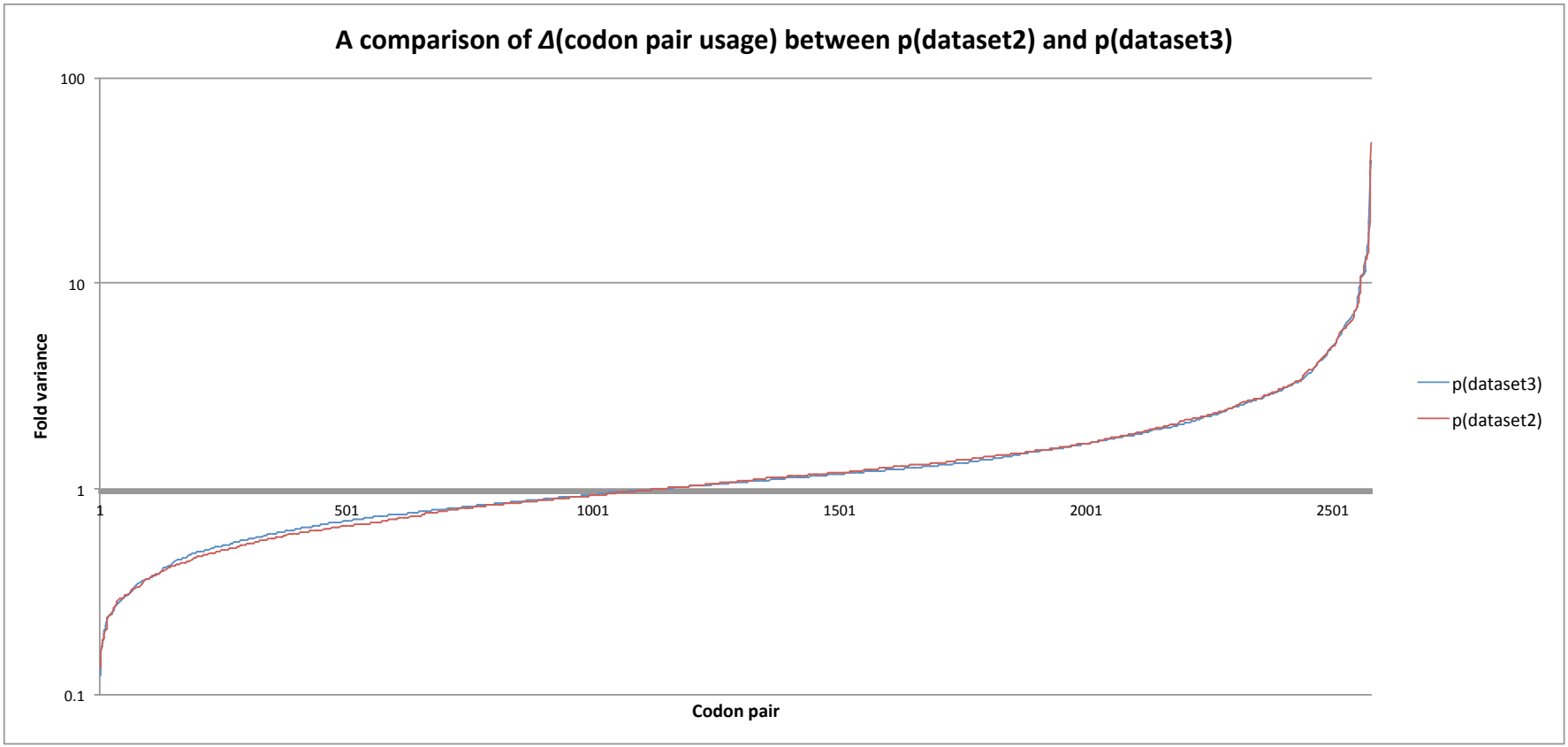


Chart 5.6 – A comparison of *Δ*(codon pair usage) between *p(dataset2)* and *p(dataset3)* shows a minor reduction in variance on amino acid pair correction

A comparison between *p(dataset2)* and *p(dataset3)* variance against observed codon pairs from *o(dataset1)* shows only a very minor decrease in variation for *p(dataset3),* which still shows considerable variation from the observed codon pair distribution.

5.2.3 Hypothesis Three – Global transgene codon usage influences absolute gene expression

Global (in this case referring to gene-wide) codon usage can be quantified in the form of a Codon Adaptation Index (CAI) as discussed above (Sharp and Li, 1987). The general consensus is that a higher CAI should correspond with higher expression. However, this has yet to be rigorously investigated in *C. reinhardtii*.

A panel of transgenes previously transformed into the *C. reinhardtii* chloroplast by members of the Purton lab were investigated. The panel consisted of eleven expressing and five non-expressing genes and were analysed using the CUO sub program, 'CAI analyser'. As in Hypotheses 1 and 2, genes were analysed relative to the complete 68 gene set rather than a high expressing sub-set. This was following the logic that as all chloroplast genes are successfully expressed, a complete complement codon usage table should be capable of detecting lethal as opposed to non-optimal gene features.

The panel of 16 recombinant genes, their expression status, and CAI scores are displayed in Table 5.2 and Chart 5.7. Both expressed and non-expressed genes show a full range of CAI values, indicating that a global codon usage-based CAI score alone is not sufficient to detect features that could block gene expression.

Table 5.2 – A test set of eleven expressed and five non-expressed transgenes, as transformed into the *C. reinhardtii* chloroplast by members of the Purton lab

All genes were optimised and synthesised by GeneArt unless otherwise stated

Gene	CAI	Pair CpAI	Expressed
lacI	0.993	0.891	Yes
cpl-1	0.985	0.88	Yes
hGH	0.97	0.833	Yes
codA (CUO optimised)	0.867	0.745	Yes
Pal	0.864	0.705	Yes
phi-11	0.781	0.566	Yes
Spy	0.759	0.555	Yes
ereB (E. coli)	0.464	0.205	Yes
arg4 (S. cerevisiae)	0.378	0.134	Yes
argH (Synechocystis)	0.26	0.074	Yes
codA (E. coli)	0.209	0.069	Yes
codA (Synechocystis)	0.29	0.085	No
codA (S. cerevisiae)	0.355	0.12	No
gp20	0.772	0.491	No
lys16	0.821	0.662	No
shbp (CUO optimised)	0.989	0.903	No

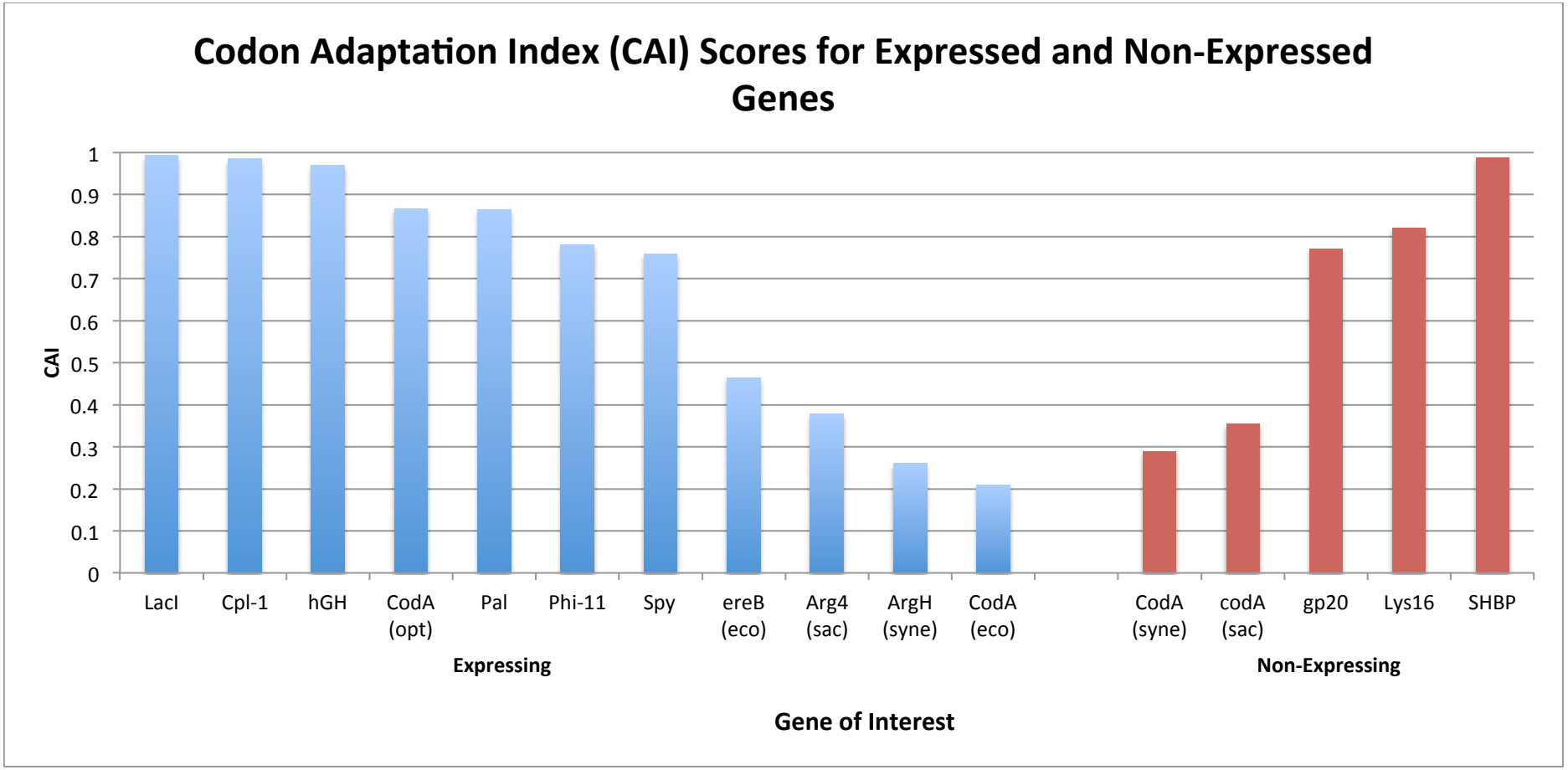


Chart 5.7 – Wide ranges of Codon Adaptation Index (CAI) scores are observed for both expressing and non-expressing transgenes

Global codon usage scores in the form of Codon Adaptation Indices for the eleven expressing (blue) and five non-expressing (red) transgenes. From these data, it seems apparent that there is no correlation between global codon usage and absolute protein expression.

5.2.4 Hypothesis Four – Global transgene codon pair usage influences absolute gene expression

It has been shown above that global codon usage alone cannot be used to discern between expressing and non-expressing genes; however, a standard CAI does not take codon pairing into account. In order to rectify this, a new module was added to the CUO (included in beta 0.91) to produce a novel quantitative value, the Codon pair Adaptation Index (CpAI). The same panel of 16 genes was analysed as for Hypothesis three; CpAI data are displayed in Table 5.2 and Chart 5.8.

Again, a wide range of values was seen in both expressing and non-expressing subsets, implying that global codon pair use is also not sufficient in this case to predict expression of transgenes, although such measures may prove important in determining levels of expression.

Given the data presented in the preceding chapter on endolysin chimeras and full fusion proteins, it seems likely that non-expression in such cases could be due to stalling of the ribosome during elongation. Such an event would point more towards local codon or codon pairing issues rather than a global effect. This is investigated in Hypothesis five below.

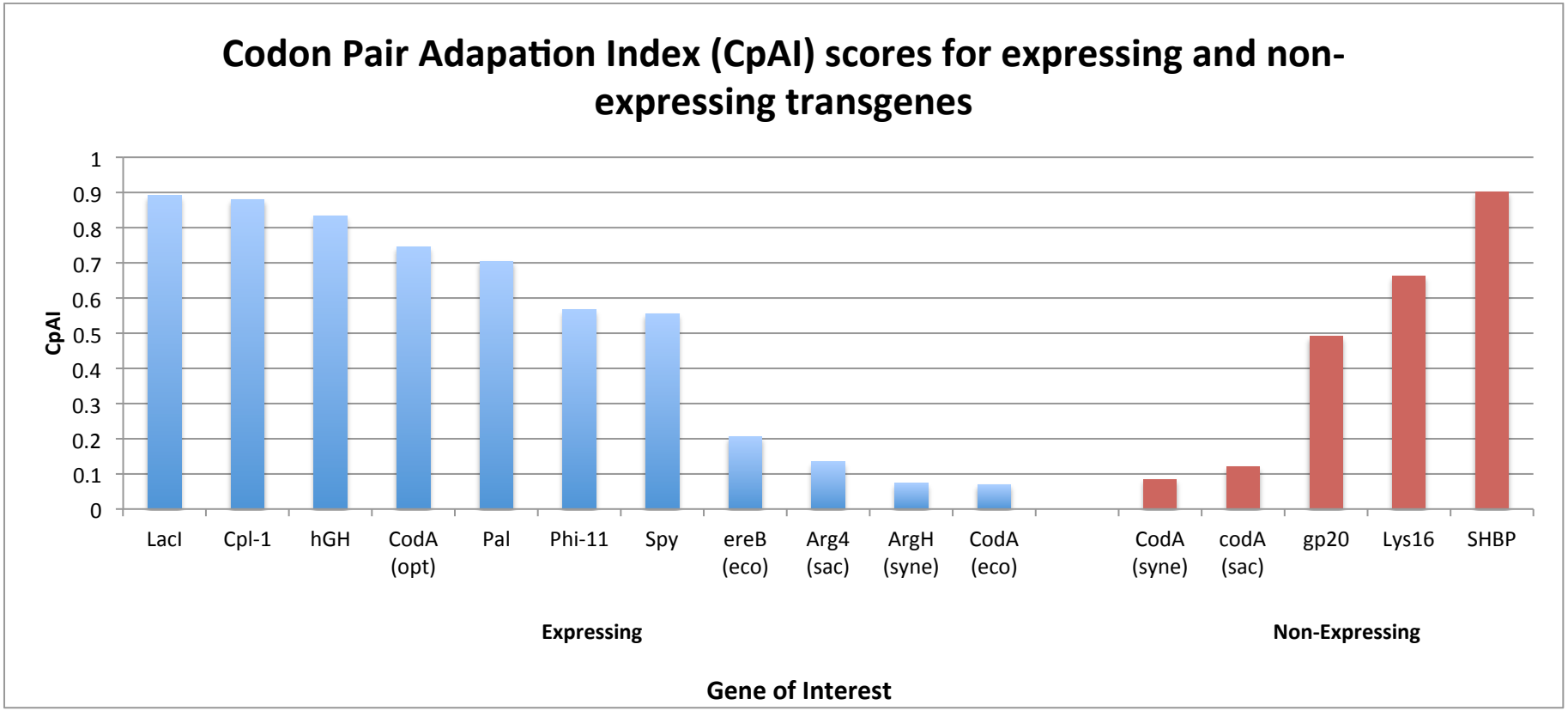


Chart 5.8 – A wide range of Codon pair Adaptation Index (CpAI) scores are observed for both expressing and non-expressing transgenes

Global codon pair usage is displayed in the form of Codon pair Adaptation Indices for the eleven expressing (blue) and five non-expressing (red) transgenes. In respect of individual CpAI scores, there is a wide range of CpAI scores seen for both expressing and non-expressing genes, indicating that CpAI cannot be directly linked to absolute expression.

5.2.5 Hypothesis Five – Local codon pair usage influences recombinant gene expression

It has been shown by Weiβ and colleagues that in the *C. reinhardtii psbA* gene, when the rare CGG and AGG codons are introduced into a region encoding a loop of the D1 protein of photosystem II, they are sufficient to block levels of expression required to allow phototrophic growth (Weiß *et al.*, 2012). It has yet to be investigated whether there are rare codon pairs that could have a similar effect, thus this hypothesis focuses on the unseen codon pairs found in *o(dataset1)*, and their relation to recombinant genes expressed in the Purton lab. Initial work on rare local codon pair interactions was, for simplicity, based purely around codons not seen at all in the *C. reinhardtii* chloroplast genome. There are 1142 such zero scoring codon pairs (ZSCPs) observed, making up 30.7 % of all non stop-containing codon pairs. The complete absence of these pairs prompts the question: are they being avoided because they cannot be tolerated, or is their absence simply a consequence of low probability of occurrence and a small genome? The former possibility is investigated in this section, with the latter being explored in Hypothesis Six.

To address this question, the codon pair usage in the above panel of transgenes (Table 5.2) was assessed for zero scoring codon count. Direct analysis of the codon pairs used in these genes was made possible due to another addition to CUO beta 0.91 – the ability to export codon pair data to the clipboard, and then into Microsoft Excel. Once in Excel, the total number of zero scoring codons was counted using the =CountIf command.

The results of this analysis are shown in Chart 5.9. It can be seen that ZSCPs are, in fact, seen on numerous occasions in both expressing and non-expressing transgene subsets. The majority of these are seen in heterologous genes, although there are also several observed in codon-optimised synthetic genes, as designed by GeneArt. These data confirm that at least some ZSCPs are completely tolerated by the *C. reinhardtii* chloroplast expression machinery, even at high frequencies.

As if to further verify the lack of effect that ZSCPs seem to have, of all the genes examined only a non-expressed transgene, Lys16, is completely devoid of zero

scoring codons. That is to say, that all the codon pairs seen in the Lys16 coding sequence are used at least once in the native *C. reinhardtii* chloroplast genome. An important point to note here, however, is that the occurrence of a codon pair in the native context does not necessarily mean that it is tolerated in all contexts. As shown by Weiβ *et al.*, rare but not unseen arginine codons were sufficient to block *psbA* expression to levels required for selection (Weiß *et al.*, 2012). The converse is exemplified by *codA* from *E. coli*, where the rare arginine codons reported by Weiß and colleagues are seen six times each, yet the gene still shows detectable levels of expression.

The low use of zero scoring codon pairs in highly expressed genes such as Cpl-1, and extensive use in other expressed genes, shows that at least some zero scoring codon pairs are not avoided due to explicit negative effects. It is likely that many of the ZSCPs observed in expressed transgenes are not seen in the *C. reinhardtii* chloroplast genome simply as a result of the small size of the genome precluding such rare combinations from arising. This postulate is assessed in detail in Hypothesis Six.

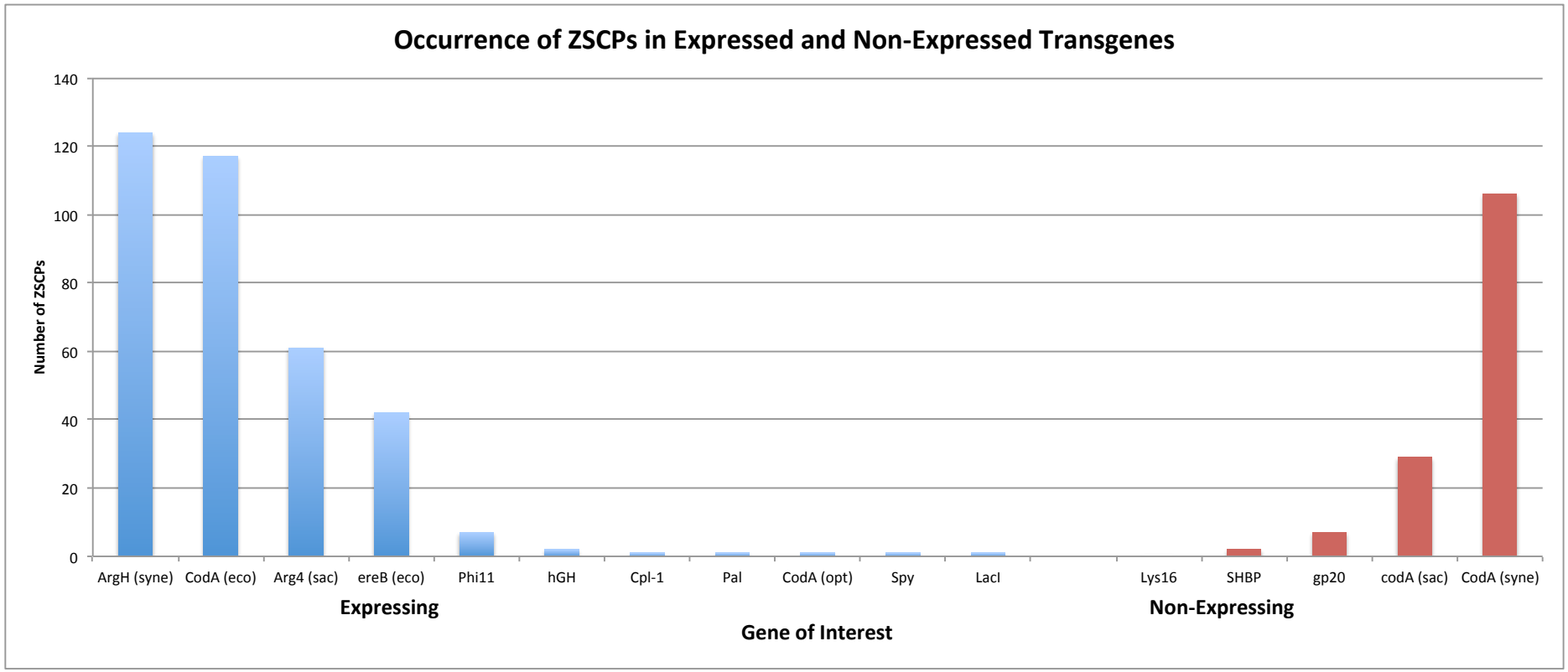


Chart 5.9 – Zero Scoring Codon Pairs are observed in both expressing and non-expressing transgenes

Codon-optimised genes show a very low occurrence of ZSCPs, whereas non-optimised genes show ZSCPs in far greater numbers. As with global codon- and codon pair usage, a wide range of ZSCP occurrences are seen for both expressed and non-expressed transgenes, suggesting little correlation between occurrence of ZSCPs and absolute expression.

5.2.6 Hypothesis Six – All zero scoring codon pairs observed for the *C. reinhardtii* chloroplast are explicitly avoided

It has been shown above (5.2.2) that codon pair bias does, at least on some level, exist in the *C. reinhardtii* chloroplast genome. It has also been observed that 1142 codon pairs are absent from the *C. reinhardtii* chloroplast genome, the so-called Zero Scoring Codon Pairs (ZSCPs). It has yet to be investigated whether these codons are explicitly avoided as a function of the codon bias already demonstrated, or if, given the small genome in consideration, they are simply too rare to be expected to occur, based on their codon makeup and amino acid products. It was considered whether the absence of these codon pairs could be a consequence of extremely rare individual codons in combination with the small genome size. As a preliminary investigation, an expected figure for ZSCPs purely as a result of individual codon usage was generated based on absolute individual codon frequencies.

To assess the likelihood of a particular codon pair presenting itself in the *C. reinhardtii* chloroplast genome, the individual codons were split into two groups: High and Low. The High group was defined as any codon which, if paired with itself or a more abundant codon, would give a combined predicted occurrence of 1 or more. The Low subset was defined correspondingly as a codon which, if paired with itself or a less abundant codon would give a combined expected value of below 1:

$$p(C_{high} : C_{high}) * 25324 \geq 1$$

$$\therefore$$

$$\sqrt{p(C_{high})} \geq 1 \,/\, 25324$$

$$\therefore$$

$$p(C_{high}) \geq 0.628\% \qquad\qquad p(C_{low}) \leq 0.628\%$$

From predicted High and Low group boundaries, predicted occurrences of High: High, High: Low, Low: High, and Low: Low (HH, HL, LH, LL) were calculated, and compared to the total number of pair combinations in each subset. By definition, you would expect to see more than one occurrence per codon pair for each HH, and consequently, less than one occurrence per codon pair for any LL pairing. HL or HL pairings could give either result.

Table 5.3 illustrates the absolute individual usage of codons across the *C. reinhardtii* chloroplast genome. The heat map shows the abundance of each codon, with green representing the most abundant, and red the least. It can be seen that a number of codons shown very low absolute frequencies, suggesting that the resulting codon pairs would have a very low occurrence, if seen at all.

Table 5.4 shows the predicted values for the High/Low codon pairings. Groups HH, LH, and HL all show predicted occurrence considerably higher than the number of possible codon combinations, making it likely that all pairings in these groups should be seen in the *C. reinhardtii* chloroplast, unless specifically avoided. The LL group, however, has a predicted value of only 142, despite there being 625 possible codon combinations in this subset. This implies that even in a situation where each of the 142 pairs seen were unique, there would still be 483 codons that would be absent from the *C. reinhardtii* chloroplast genome by statistical probability. These data suggest that in the complete absence of a codon pair bias, there are a considerable number of codon pairs that would not be expected to occur in a genome of this size, and thus are not specifically avoided.

Table 5.3 – Relative and absolute frequencies of the 61 non-stop codons seen in the *C. reinhardtii* chloroplast genome

Absolute frequencies are displayed as a heat map, with green indicating the most frequently used codons, red the least. Of the 61 codons, 30 occur in less than 1 % of cases.

Amino acid	Codon	Number	Relative Frequency	Frequency per thousand	Absolute Frequency
Lys	AAA	1940	0.94	76.4	0.076402016
Asn	AAC	468	0.28	18.43	0.018431002
Lys	AAG	122	0.06	4.8	0.004804663
Asn	AAU	1185	0.72	46.67	0.046668242
Thr	ACA	859	0.51	33.83	0.033829553
Thr	ACC	86	0.05	3.39	0.003386894
Thr	ACG	95	0.06	3.74	0.003741336
Thr	ACU	651	0.38	25.64	0.025637996
Arg	AGA	141	0.12	5.55	0.00555293
Ser	AGC	117	0.07	4.61	0.00460775
Arg	AGG	19	0.02	0.75	0.000748267
Ser	AGU	424	0.24	16.7	0.016698173
Ile	AUA	182	0.11	7.17	0.007167612
Ile	AUC	204	0.12	8.03	0.008034026
Met	AUG	540	1	21.27	0.021266541
Ile	AUU	1340	0.78	52.77	0.052772527
Gln	CAA	1016	0.93	40.01	0.040012602
His	CAC	222	0.45	8.74	0.008742911
Gln	CAG	81	0.07	3.19	0.003189981
His	CAU	271	0.55	10.67	0.010672653
Pro	CCA	576	0.55	22.68	0.02268431
Pro	CCC	43	0.04	1.69	0.001693447
Pro	CCG	53	0.05	2.09	0.002087272
Pro	CCU	380	0.36	14.97	0.014965343
Arg	CGA	91	0.08	3.58	0.003583806
Arg	CGC	66	0.06	2.6	0.002599244
Arg	CGG	6	0.01	0.24	0.000236295
Arg	CGU	862	0.73	33.95	0.0339477
Leu	CUA	176	0.06	6.93	0.006931317
Leu	CUC	14	0.01	0.55	0.000551355
Leu	CUG	47	0.02	1.85	0.001850977
Leu	CUU	377	0.14	14.85	0.014847196
Glu	GAA	1062	0.92	41.82	0.041824197
Asp	GAC	202	0.24	7.96	0.007955261
Glu	GAG	87	0.08	3.43	0.003426276
Asp	GAU	646	0.76	25.44	0.025441084
Ala	GCA	520	0.33	20.48	0.020478891
Ala	GCC	100	0.06	3.94	0.003938248
Ala	GCG	78	0.05	3.07	0.003071834
Ala	GCU	873	0.56	34.38	0.034380907
Gly	GGA	202	0.13	7.96	0.007955261
Gly	GGC	115	0.07	4.53	0.004528986
Gly	GGG	91	0.06	3.58	0.003583806
Gly	GGU	1130	0.73	44.5	0.044502205
Val	GUA	675	0.45	26.58	0.026583176
Val	GUC	19	0.01	0.75	0.000748267
Val	GUG	96	0.06	3.78	0.003780718
Val	GUU	699	0.47	27.53	0.027528355
Tyr	UAC	234	0.26	9.22	0.009215501
Tyr	UAU	668	0.74	26.31	0.026307498
Ser	UCA	594	0.34	23.39	0.023393195
Ser	UCC	57	0.03	2.24	0.002244802
Ser	UCG	98	0.06	3.86	0.003859483
Ser	UCU	477	0.27	18.79	0.018785444
Cys	UGC	20	0.1	0.79	0.00078765
Trp	UGG	326	1	12.84	0.012838689
Cys	UGU	175	0.9	6.89	0.006891934
Leu	UUA	2034	0.74	80.1	0.08010397
Phe	UUC	426	0.33	16.78	0.016776938
Leu	UUG	99	0.04	3.9	0.003898866
Phe	UUU	867	0.67	34.14	0.034144612

Table 5.4 – Specific probabilities for the Low and High codon groups, and the compound probabilities generated for the four codon pair groups

These probability values are used to predict codon pair occurrences for the *C. reinhardtii* chloroplast genome as a whole. It can be seen that, as expected, the codon pairs in the p(LL) group will be seen considerably less than the total number of codons in that group. As such, at least 483 codon pairs are not expected to occur.

Event	Probability		
p(L)	0.075		
p(H)	0.925		
		Codon pairs	Predicted for whole genome
p(LL)	0.006	625	142
p(HH)	0.856	784	21668
p(LH)	0.069	924	1757
p(HL)	0.069	924	1757

Once it was demonstrated that there was indeed a subset of codon pairs that would be unlikely to occur in a genome of this size, a more in-depth investigation into ZSCPs was instigated. For this study, the more accurate *p(dataset3)* was utilised. Expected values were tabulated and all codon pairs with a predicted value of less than 1 were counted to give a predicted ZSCP value. Interestingly, for this model the total number of codon pairs with a predicted value of less than 1 was 1364, 222 higher than the number actually observed. Taken in isolation, this seems to be enough evidence to discard the ZSCPs seen in the *C. reinhardtii* chloroplast genome as a statistical artefact. However, closer scrutiny of the data reveals that the predicted Zero Scoring Codon pairs, pZSCPs, do not necessarily match the observed ZSCPs. By filtering *p(dateset3)* by those codon pairs not observed in the *C. reinhardtii* chloroplast genome, it becomes apparent that of the 1142 observed ZSCPs, 189 are actually predicted to be present by this model, implying that in some cases, codon pairs are explicitly avoided. These codons pairs are shown in Chart 5.10 against their predicted value. These ZSCPs with predicted values greater than 1 as modelled by *p(dataset3)* are henceforth referred to unexpectedly unseen Zero Scoring Codon Pairs (uuZSCPs).

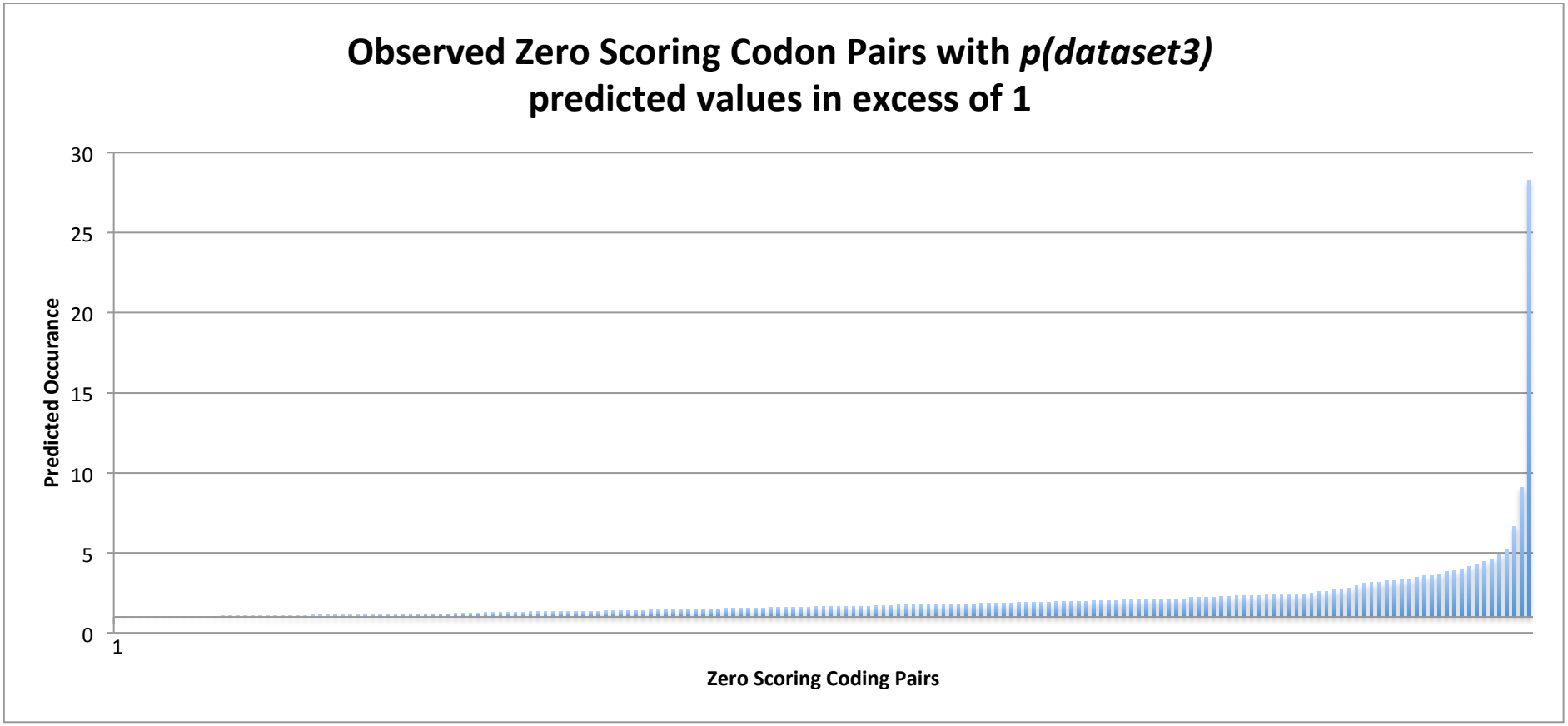


Chart 5.10 – Observed Zero Scoring Codon Pairs with *p(dataset3)* predicted values in excess of 1

Of the 1142 ZSCPs seen for the *C. reinhardtii* chloroplast genome, 953 are not predicted to occur in a genome of this size, based on the *p(dataset3)* model. The remaining 189 of the observed ZSCPs are predicted to occur, but do not. Many of these have predicted values barely in excess of 1, although several have considerably higher values.

It can be seen that the majority of these codons are predicted at low frequencies, with 70 % having predicted values below 2, and 88 % below 3. Some, though, are predicted at far higher values. Of particular interest are GAU:CUU and UUU:CGU, encoding Asp:Leu and Phe:Arg respectively (shown in table 4). The degree to which these codons are over-represented based on amino acid and relative codon usage suggests an explicit avoidance, perhaps due to a particularly unfavourable steric interaction.

To return to the initial hypothesis – going by the predicted values from *p(dataset3)* in the absence of a codon bias, it is predicted that slightly more ZSCPs should occur than actually do. The apparent over-representation of some codon pairs suggests another interesting area of study, although is beyond the remit of this investigation. The set of ZSCPs that is observed does not entirely align with the predicted ZSCPs, with some codon pairs significantly under-represented in the *C. reinhardtii* chloroplast genome. The extent to which some pairs are avoided suggests a specific bias against their use. Whether or not this is reflected in recombinant genes is explored in Hypothesis Seven.

Table 5.5 – Two ZSCPs show particularly high predicted values as generated by *p(dataset3),* suggesting explicit avoidance

The *p(dataset3)* model, where no specific codon pair bias is featured, states that these codon pairs should be observed due to their relatively high individual codon usages and frequent amino acid pair occurrences. Their absence from the observed dataset *o(dataset1)* strongly suggests an unknown factor actively blocking their use.

Amino Acid Pair	First Codon	Relative Frequency 1	Second Codon	Relative Frequency 2	Combined Relative Frequency	Total Observed for Amino Acid Pair	**Predicted for Pair**	**Actual Occurrence**
Phe:Arg	UUU	0.67	CGU	0.73	0.49	58.00	**28.29**	**0**
Asp:Leu	GAU	0.76	CUU	0.14	0.10	87.00	**9.10**	**0**

5.2.7 Hypothesis Seven – The failure of non-expressing genes can be explained by the presence of uuZSCPs

Hypothesis six has suggested that unexpectedly unseen zero scoring codon pairs (uuZSCPs) are explicitly avoided by the *C. reinhardtii* chloroplast, possibly due to deleterious effects on gene expression. It was therefore deemed useful to investigate whether any of the recombinant genes so far transformed into the *C. reinhardtii* chloroplast in the Purton lab contained any of these codon pairs.

In order to focus on codon pairs most strongly avoided (and hence those most likely to have a negative effect), only uuZSCPs with predicted values in excess of four were investigated, totalling nine pairs. Codon pair datasets were generated for the panel of transgenes described in Hypothesis three, hereafter referred to as *e(dataset1)* and *n(dataset1)* for expressing and non-expressing genes, respectively. These datasets were cross-referenced against the nine uuZSCPs selected (Table 5.6). From these data it can be seen that, contrary to expectations, these codon pairs are fairly well represented in both datasets. Of the nine uuZSCPs, six (including the two most avoided in the *C. reinhardtii* chloroplast genome), are present in expressing transgenes. Also of note is that there are no cases where a uuZSCP is present in a non-expressing gene but not an expressing one, thus allowing a direct comparison. This is likely a consequence of the limited size of the non-expressed dataset.

Table 5.6 – The nine uuZSCPs most avoided in the *C. reinhardtii* chloroplast genome in relation to the previously examined panel of expressing and non-expressing transgenes

Despite the apparent avoidance of these codons in the *C. reinhardtii* chloroplast genome, these data show that at least some are tolerated in recombinant genes. There are no cases where a uuZSCP is seen in a non-expressing but not an expressing gene, such to suggest a possible negative impact of uuZSCPs on transgene expression. Analysis of more transgenes genes will be required for a more definitive picture.

Codon pair	First codon	Second codon	Predicted occurrence	Recombinant expressing (codon pairs/ αα pair total)	Recombinant non-expressing (codon pairs/ αα pair total)
Phe:Arg	UUU	CGU	28.2	4 of 12	1 of 1
Asp:Leu	GAU	CUU	9.10	2 of 28	0 of 10
Phe:His	UUU	CAC	6.64	0 of 0	0 of 0
Lys:Thr	AAG	ACA	5.23	0 of 16	0 of 3
Ser:His	UCU	CAU	4.90	3 of 10	1 of 1
Gly:Leu	GGU	UUG	4.61	3 of 30	0 of 4
Ile:Leu	AUU	UUG	4.45	2 of 17	1 of 3
Leu:Ser	CUA	UCA	4.31	1 of 20	0 of 10
Thr:Gly	ACG	GGU	4.13	0 of 18	0 of 6

The presence of these uuZSCPs in successfully expressed transgenes, and their relative absence from the non-expressed subset shows that, although uuZSCPs are apparently avoided in the *C. reinhardtii* chloroplast genome, they are not inherently lethal to expression. Comment cannot be made about those codon pairs not seen in either expressed or non-expressed datasets, although this is likely to be addressed as more transgenes are transformed into the *C. reinhardtii* chloroplast in the Purton lab. It should also be recognised that, especially in the case of the non-expressed genes, the sample sizes are very small, and more data will be required to give a reliable picture.

Another factor, as of yet intentionally ignored, is the effect that codon pairing has on levels of expression. The issue of recombinant protein yield has been avoided for the reasons stated at the start of this chapter; however, it is worth acknowledging that the most strongly avoided codon pair in the *C. reinhardtii* chloroplast genome, UUU:CGU, is present in one of the most highly expressing recombinant genes observed in the Purton lab, *cpl-1*. This could either imply that even the most avoided uuZSCPs do not negatively affect expression, and thus the concern regarding extreme codon pair use should be abandoned; or, alternatively, that the high expression is seen *despite* the presence of this codon. Until a suitable control such as a site-directed mutant of *cpl-1* with the removal of this codon pair is generated, it is not possible to comment on which of these is more likely to be the case.

5.2.8 Hypothesis Eight – Regions of poor codon pair usage are responsible for non-expressing recombinant genes

It has now been shown that there does indeed seem to be a codon pair bias present in the *C. reinhardtii* chloroplast genome. This bias, however, does not appear to greatly influence transgene expression, with neither global codon pair usage, individual codon pair usage, nor incorporation of specifically avoided codon pairs seeming to correlate with absolute expression of transgenes. Another possibility is that in order to sufficiently destabilise the ribosome:tRNA:mRNA complex and stall elongation, a succession of 'bad' codon pairs is required. This hypothesis seeks to investigate if regions of poor codon pair usage are seen more frequently in non-expressed recombinant genes relative to their expressed counterparts. To analyse local regions of poor codon pair usage in recombinant genes, codon pair adaptation data was generated by the CUO and exported into Microsoft Excel for analysis. This data consisted of the weightings of each codon pair used in ten genes, so selected to give a spread of expressed, non-expressed, optimised, and non-optimised genes as shown in Table 5.7.

Table 5.7 – Recombinant genes chosen for local region codon pair analysis

The selected genes were nominated so to give as diverse a spread of native, GeneArt-optimised, and CUO-optimised genes as possible.

	Expressed	**Non-Expressed**
Optimised	*cpl-1* *pal* *codA (CUO-optimised)*	*gp20* *lys16* *shbp (CUO-optimised)*
Non-Optimised	*ereB (E. coli)* *codA (E. coli)*	*codA (Synechocystis)* *codA (S. cerevisiae)*

The codon pair adaptation for the above genes was plotted as a function of position on the gene. In order to more clearly visualise regions of poor codon pair usage as opposed to single point values, a moving average was also plotted. A six-codon pair window was chosen so to allow two sets of three non-overlapping pairs, thus incorporating sufficient up- and down-stream sequence to mitigate isolated bad codon events. The results of these analyses are shown in Chart 5.11-Chart 5.14. For clarity, the optimised and non-optimised subsets will be considered separately.

There is considerable variation in codon pair usage for the optimised group of genes, even within those which do/ do not express. Some genes do suggest a loose correlation between regions of poor codon usage and expression: *cpl-1, pal* and *codA* (optimised) are all expressed, and show a consistently higher codon pair adaptation than *gp20* and *lys16*, which are not. The latter two genes frequently display moving averages of below 0.4 (seen only once in the expressed subset), and *gp20* shows averages below 0.2 on two occasions.

Contradicting these tentative summations, however, is the final non-expressed optimised gene, *shbp*, which shows consistently high codon pair usage, exceeding that of each of the expressed genes. The *shbp* gene does show two cases of ZSCPs, but on closer evaluation neither are significant. One, UCA:UGU, is seen in two other expressed recombinant genes, and the other, GGA:UCC, has a predicted occurrence in the *C. reinhardtii* chloroplast genome of 0.31; thus it is not specially avoided, reducing the likelihood of it being detrimental.

If the optimised genes evaluated present some contradiction as to possible patterns, the non-expressed subset is much clearer – both the expressed and non-expressed groups have extremely low codon pair usage throughout. The lowest codon pair adaptation seen is, in fact, from the expressed gene, *codA* (*E. coli*), where large portions of the moving average are below 0.2.

The data available suggests that there is no correlation between the absolute values of a 6-point moving average of codon usage and transgene expression. One point to note, however, is in relation to consistency. Although much of the expressed *codA* (*E. coli*) gene displays very poor codon pair usage, it is regularly

punctuated by 'good' codon pairs. In the non-expressed *codA* (*Synechocystis)* gene, the overall codon pair usage is generally better than its expressed counterpart, except for one region between 195 and 240 codon pairs, where the moving average does not rise above 0.2. This region of particularly low codon pair adaptation is most clearly observed in Chart 5.15, which shows a broader 20-point moving average of both the expressed *codA* (*E. coli*), and the non-expressed *codA* (*Synechocystis)*.

It is conceivable that such a stretch of bad codon pairing could be sufficient to disrupt translation; however, without further data, this is merely speculation at this juncture. It is clear that even if the run of bad codon pairs was enough to disrupt expression in this case, it was not the causative factor in, for example, *shbp.* On a practical level, this result is also less useful – *codA (Synechocystis)* represents an extreme case, and one unlikely to arise in a typical gene design situation, unless a region of specifically bad codon usage was intentionally included.

As has become clear throughout this chapter, the issue of non-expressing transgenes in the *C. reinhardtii* chloroplast is not a simple on/off situation. It is likely that many factors contribute towards the expression (or lack of expression) of a particular gene. Codon- and codon pair usage are likely to be involved in this process; however, this investigation has yet to reveal any cases where they are the single deciding factor.

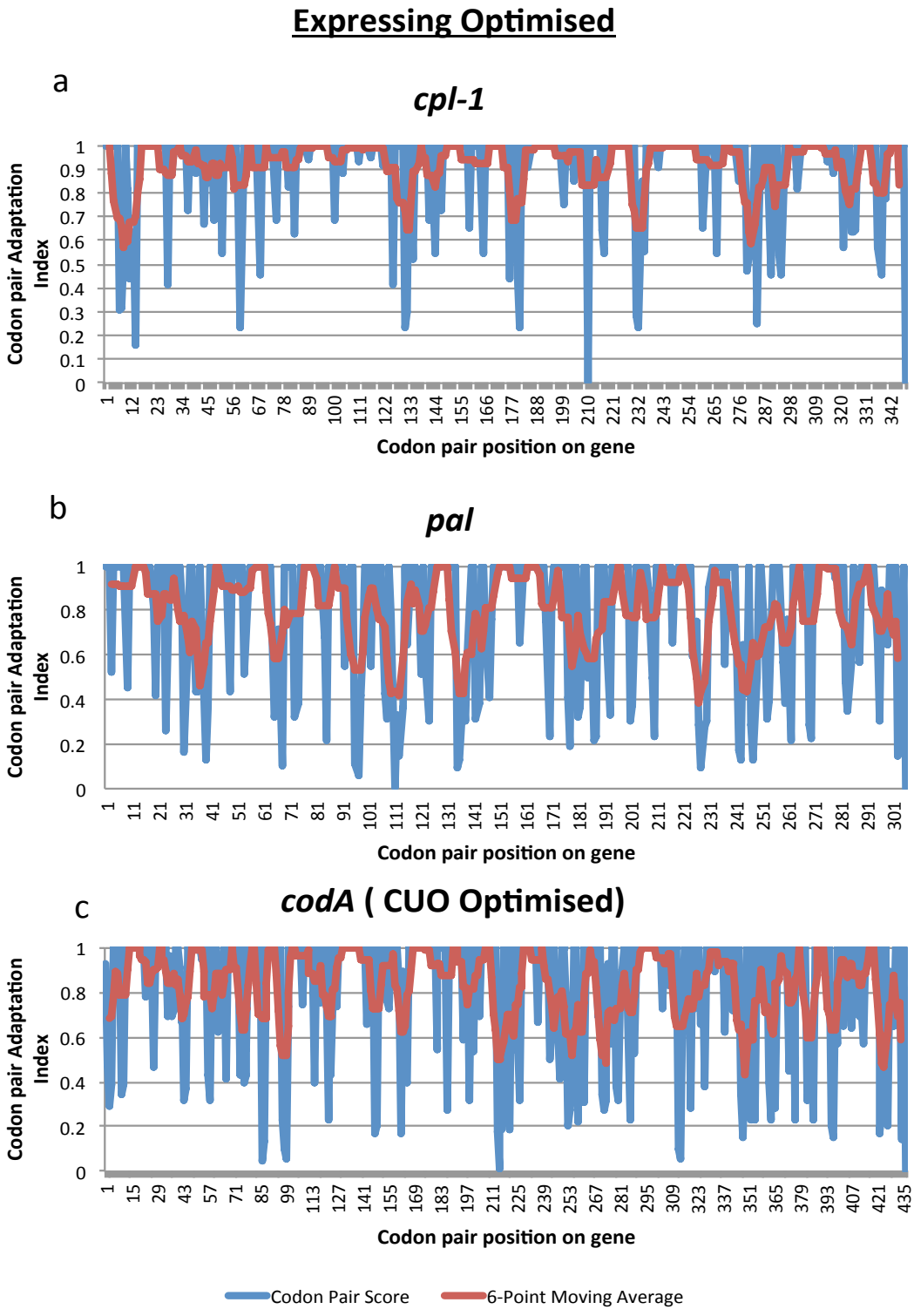


Chart 5.11 – Local regional codon pair analysis of three expressed optimised transgenes: a) *cpl-1*, b) *pal*, and c) *codA* (CUO optimised)

The codon pair use in the above three genes is generally fairly good, as shown by the 6-point moving average which rarely falls below a codon pair adaptation score of 0.6. All three genes show isolated bad codon pairs, although this is not seen to block absolute expression.

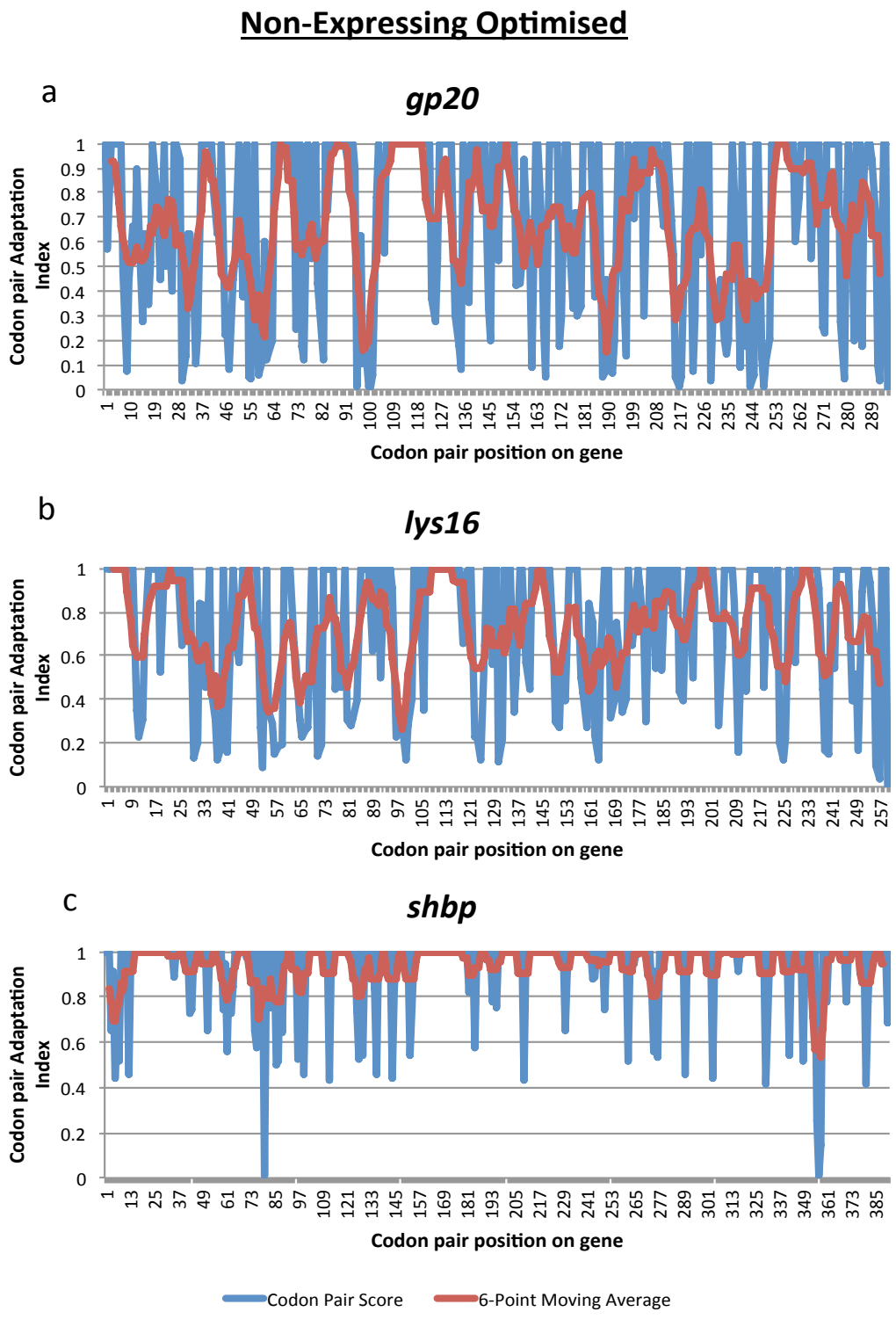


Chart 5.12 – Local regional codon pair analysis of three non-expressing optimised transgenes: a) *gp20*, b) *lys16* and c) *shbp*

The codon pair use in the above three genes is shown to be considerably less favourable than that seen for the optimised genes above, with the exception of *shbp*, which is substantially better.

Expressing Non-Optimised

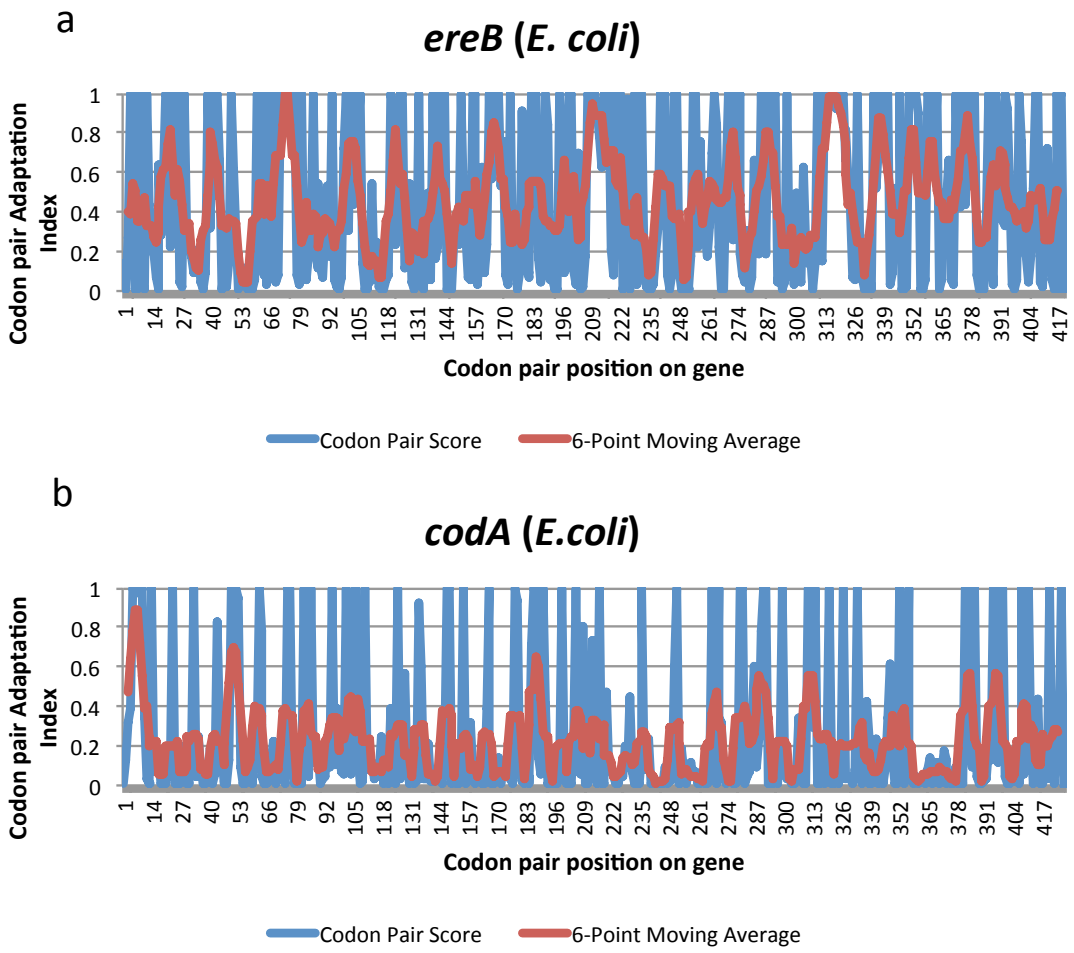


Chart 5.13 – Local regional codon pair analysis of two expressing non-optimised transgenes: a) *ereB (E. coli)* and b) *codA (E. coli)*

Both genes shown here display extremely unfavourable codon pair usage throughout their length; however, they still have been shown to accumulate recombinant product, thus casting doubt on a link between codon pair usage and absolute expression.

Non-Expressing Non-Optimised

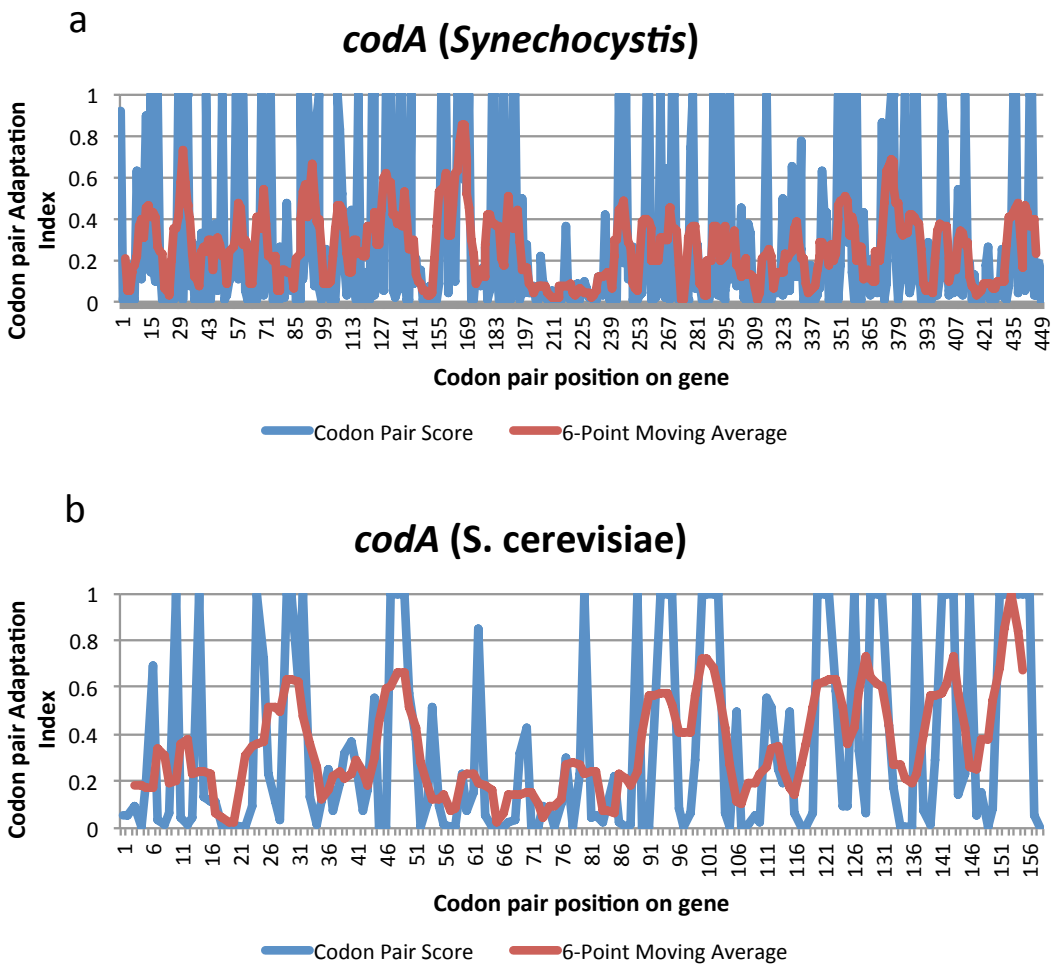


Chart 5.14 – Local regional codon pair analysis of two non-expressing non-optimised transgenes: a) *codA (Synechocystis)* and b) *codA (S. cerevisiae)*

As with the expressing non-optimised genes, these genes show very poor codon pair adaptation throughout. The only particularly apparent difference between the expressed and non-expressed non-optimised genes is not the presence of 'bad' codon pairs, but more the absence of 'good' codon pairs seen in the 195-240 region of *codA (Synechocystis).*

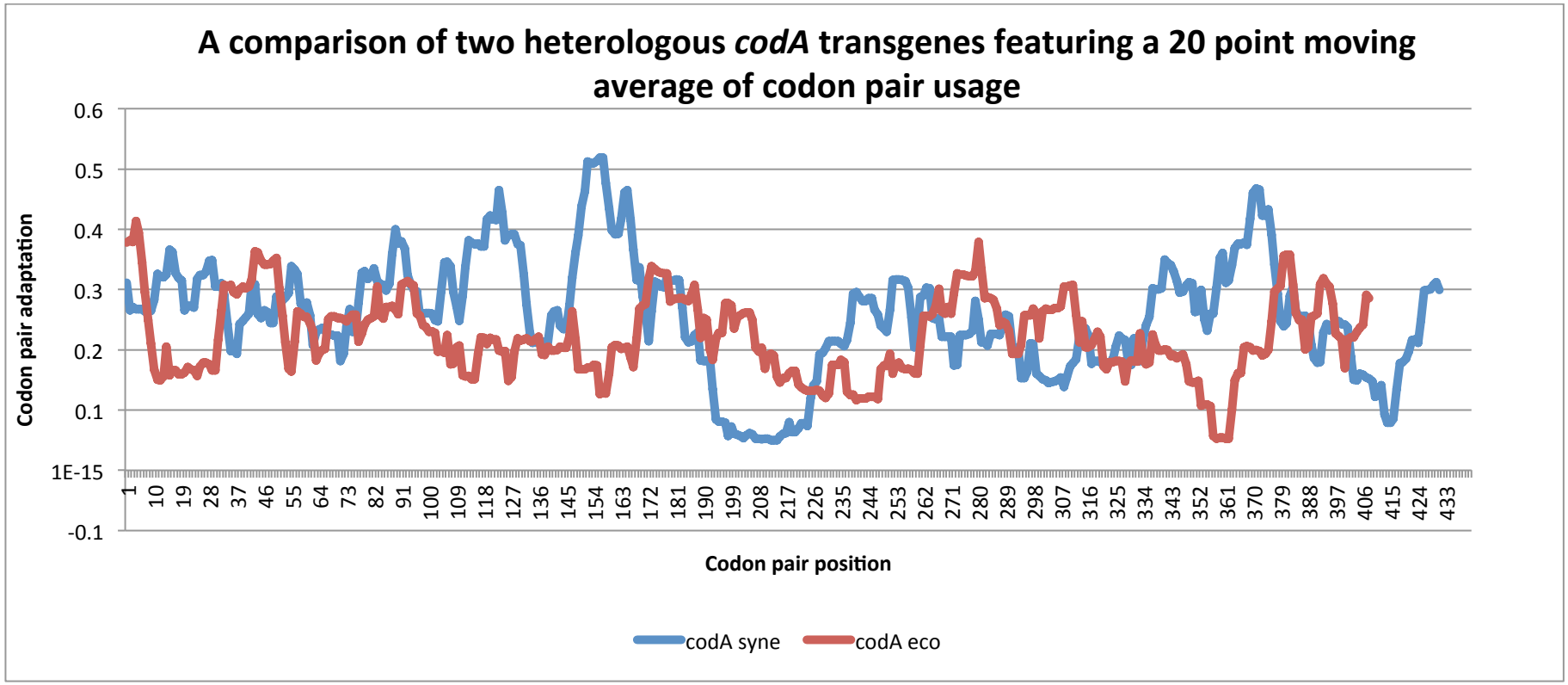


Chart 5.15 – A direct comparison of transgenic *codA* from *E. coli* (expressing) and *Synechocystis* (non-expressing), plotting regional codon pair usage using a 20-point moving average against position on the gene

The clearest difference between the expressed *E. coli* and non-expressed *Synechocystis* variants of *codA* can be seen to be the consistently bad codon usage in the non-expressing gene between positions 190 and 220. It is possible that this run of bad codons could cause ribosome stalling; however, a more detailed study is required to draw any firm conclusions.

5.2.9 Hypothesis Nine – uuZSCPs are conserved across a panel of related green algal chloroplast genomes

The causative factor of unexpectedly unseen Zero Scoring Codon Pairs (uuZSCPs) is thought to be the specific features of the environment from which they originate. A question worth addressing is thus whether these preferences are conserved in the chloroplast of related green algal species. To investigate this possibility, 14 species were analysed, and all conserved ZSCPs compared to their predicted values for the *C. reinhardtii* chloroplast genome to establish if there are uuZSCPs, at least for *C. reinhardtii*.

A panel of related, but distinct, algal species was prepared and the corresponding chloroplast genome sequences collected from Genbank (Table 5.8). These data were pooled into a single dataset which was then challenged for ZSCPs. Conserved ZSCPs were analysed using the amino acid weighted predicted values for the *C. reinhardtii* chloroplast genome (*p(dataset3*). Any codon pairs with a predicted value of less than one would hence be concluded to be non-significant in *C. reinhardtii*, thus breaking the conservation across all 14 species.

Five codon pairs were found to be absent in all 14 species, but analysis of these codon pairs in relation to their predicted values from *p(dataset3)* (Table 5.9) shows these to be extremely rare codon pairs that would not be expected to occur in the *C. reinhardtii* chloroplast genome. Clearly, without generating predicted datasets for the 13 other species, it cannot be assumed that there is not active conservation in the other members of the panel; however, this is unlikely and beyond the scope of this investigation.

Table 5.8 – 14 related green algal species selected for investigation into the conservation of Zero Scoring Codon Pairs (ZSCPs), together with their Genbank locations

The panel was selected to encompass a variety of classes while limiting the investigation to the Chlorophyta phylum (green algae)

Species	Phylum:	Class:	Order:	Family:	Genus:	NCBI Identifier
Chlamydomonas reinhardtii	*Chlorophyta*	*Chlorophyceae*	*Chlamydomonadales*	*Chlamydomonadaceae*	*Chlamydomonas*	
Dunaliella salina	*Chlorophyta*	*Chlorophyceae*	*Volvocales*	*Dunaliellaceae*	*Dunaliella*	108773031
Scenedesmus obliquus	*Chlorophyta*	*Chlorophyceae*	*Chlorococcales*	*Scenedesmaceae*	*Scenedesmus*	383930345
Pedinomonas minor	*Chlorophyta*	*Pedinophyceae*	*Pedinomonadales*	*Pedinomonadaceae*	*Pedinomonas*	376403573
Micromonas pusilla CCMP1545	*Chlorophyta*	*Prasinophyceae*	*Mamiellales*	*Mamiellaceae*	*Micromonas*	226968543
Micromonas sp. rcc299	*Chlorophyta*	*Prasinophyceae*	*Mamiellales*	*Mamiellaceae*	*Micromonas*	226968641
Pycnococcus provasolii	*Chlorophyta*	*Prasinophyceae*	*Pseudoscourfieldiales*	*Pycnococcaceae*	*Pycnococcus*	224179399
Pyramimonas parkeae	*Chlorophyta*	*Prasinophyceae*	*Pyramimonadales*		*Pyramimonas*	224179399
Chlorella vulgaris	*Chlorophyta*	*Trebouxiophyceae*	*Chlorellales*	*Chlorellaceae*	*Chlorella*	7524759
Chlorella variabilis	*Chlorophyta*	*Trebouxiophyceae*	*Chlorellales*	*Chlorellaceae*	*Chlorella*	254798615
Parachlorella kessleri	*Chlorophyta*	*Trebouxiophyceae*	*Chlorellales*		*Parachlorella*	108796875
Coccomyxa sp.C-169	*Chlorophyta*	*Trebouxiophyceae*		*Coccomyxaceae*	*Coccomyxa*	323149147
Oltmannsiellopsis viridis	*Chlorophyta*	*Ulvophyceae*			*Oltmannsiellopsis*	108773302
Pseudendoclonium akinetum	*Chlorophyta*	*Ulvophyceae*	*Ulvales*	*Kornmanniaceae*	*Pseudendoclonium*	331268093

Table 5.9 – Conserved ZSCPs across 14 related algal species chloroplasts, together with their predicted values for the *C. reinhardtii* chloroplast genome using *p(dataset3)*

The data show that for the *C. reinhardtii* chloroplast genome, the five conserved codon pairs are not uuZSCPs as their predicted values for the *C. reinhardtii* chloroplast genome are below 1, and thus do not show explicit avoidance. These predictive data are only applicable to the one species in question; however, it is sufficient to break the conservation, irrespective of whether these codons are uuZSCPs, in the other 13 species.

Amino Acid Pair	First Codon	Relative Frequency 1	Second Codon	Relative Frequency 2	Combined Relative Frequency	Total Observed for Amino Acid	Predicted for Pair
Pro:Arg	CCC	0.041	AGG	0.016	0.001	55	0.036
Cys:Arg	UGC	0.103	CGG	0.005	0.001	7	0.004
Cys:Arg	UGC	0.103	AGG	0.016	0.002	7	0.012
Arg:Arg	CGG	0.005	CGG	0.005	0.000	67	0.002
Arg:Cys	AGG	0.016	UGU	0.897	0.014	8	0.115

5.2.10 Hypothesis Ten – The available tRNA pool reflects the relative codon preferences seen in the *C. reinhardtii* chloroplast genome

Studies in other organisms, most prominently *E. coli,* have shown codon usage to be linked to both the relative levels of the corresponding tRNA, and its specific recharge rate (Plotkin and Kudla, 2011). In the *C. reinhardtii* chloroplast this is not the case, as the tRNA pool is already so restricted. All 20 amino acids are catered for; however, only 34 unique tRNAs are encoded (see Figure 5.1). As there is no evidence for tRNA import into the chloroplast, the remaining 27 codons are thought to be satisfied by virtue of the 'wobble hypothesis', as seen for many other chloroplast genomes (Marechal-Drouard *et al.*, 1993). It was initially assumed that codon preferences observed in the *C. reinhardtii* chloroplast would be primarily biased towards the tRNAs present.

To investigate whether this was the case, the tRNA genes present in the *C. reinhardtii* chloroplast genome were acquired from the literature (Maul *et al.*, 2002), and an erroneously labelled Arg tRNA corrected. The tRNA anticodons were then reverse complemented and the corresponding codons matched to the relevant relative frequencies previously generated for *p(dataset3)*. Native tRNA codons are used on average 53 % of the time, with wobble factor codons being used in the remaining 47 % of cases; however, there is a large standard deviation of 31 %. This wide variation is illustrated in Chart 5.16, which shows the proportion of codons for each amino acid catered for directly (blue) relative to the proportion that rely on the wobble effect (red). Some amino acids are shown to be encoded almost exclusively by their tRNA specific codon, for example Gln and Glu. For others, such as Cys and Ile, the inverse is true. At present, it is not clear why this might be. A possible explanation is that the wobble factor requires such preferences to be observed so as to avoid interference from other tRNAs. Given the extent of the native spread in codon preferences in regard to available tRNAs, tailoring any sort of gene design protocol based entirely on the tRNAs present in the *C. reinhardtii* chloroplast would be unwise.

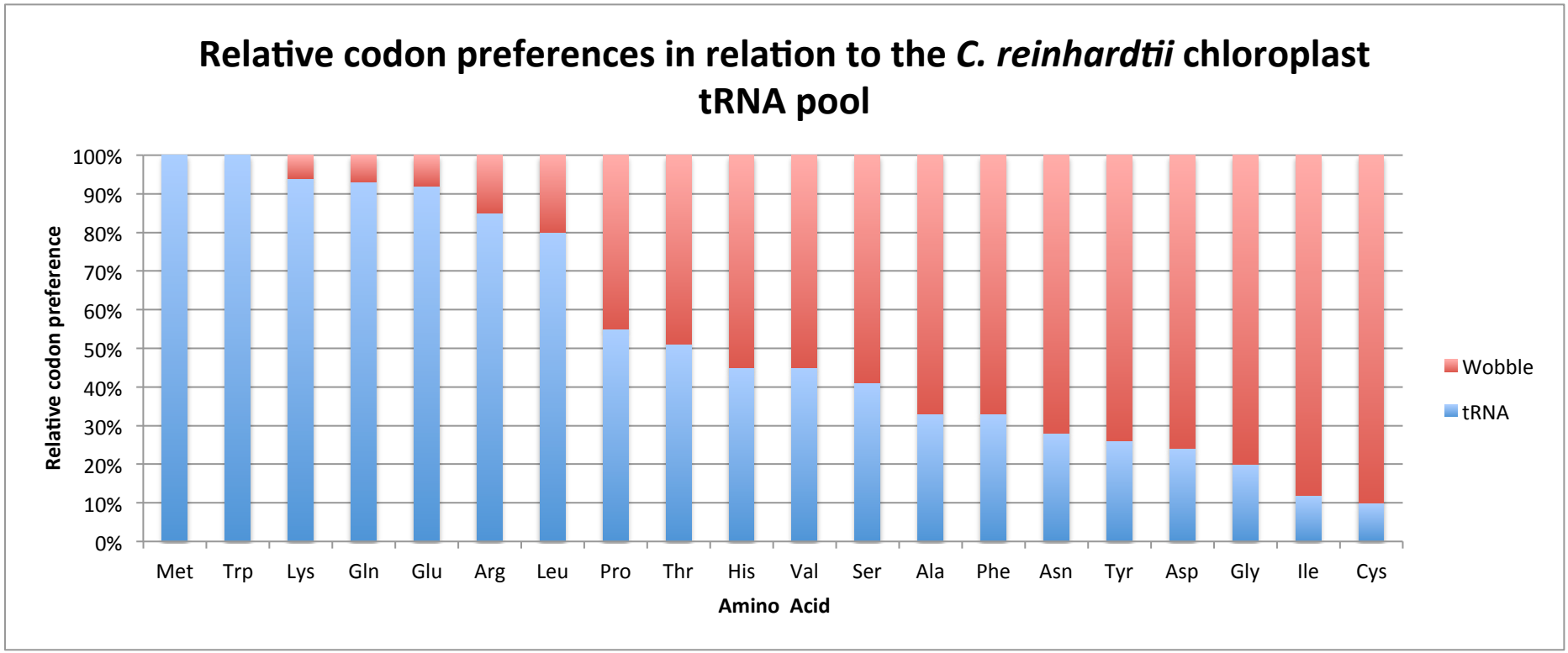


Chart 5.16 – Proportional divisions in codon use between native and wobble factor tRNAs for each amino acid in the *C. reinhardtii* chloroplast

Disregarding methionine and tryptophan, which are only encoded by one codon, all amino acids are encoded at least in part by wobble factor codons. The almost complete range of tRNA/ wobble distributions observed suggests no universal preference for either mode of recognition.

5.3 Discussion

During the course of this chapter, the relative effects of codon- and codon pair use in the *C. reinhardtii* chloroplast genome have been analysed, and the results of these analyses applied to a collection of transgenic genes in an effort to ascertain why some transgenes were successfully expressed and others were not. This goal has not been achieved; however, our understanding of the mechanics of codon- and codon pair preference in the chloroplast has been advanced.

To briefly summarise this chapter's findings:

- **Hypotheses One and Two** confirmed that the *C. reinhardtii* chloroplast genome does display specific codon- and codon pair biases relative to model predictions that exclude any such preferences.
- **Hypotheses Three and Four** examined the gene-wide codon- and codon pair adaptation of a panel of expressing and non-expressing transgenes investigated in the Purton lab. It was shown that neither measure correlated with the absolute expression of the recombinant genes examined.
- **Hypothesis Five** introduced the concept of Zero Scoring Codon Pairs – codon pairs not seen in the *C. reinhardtii* chloroplast genome. This set of 1142 codon pairs was used to challenge a panel of expressing and non-expressing transgenes, and it was found that these unseen codon pairs are actually seen extensively in both expressing and non-expressing transgene settings. This demonstrated that the absence of a codon pair from the native *C. reinhardtii* chloroplast genome does not explicitly indicate a fatal intolerance to expression.
- This topic was investigated further in **Hypothesis Six**, where predictions were made as to how many ZSCPs could be expected in the *C. reinhardtii* chloroplast genome. It was found that the non codon pair biased model *p(dataset3)* predicted 1364 ZSCPs, 222 more than are actually seen. These predicted pairs were not, however, all in agreement with the observed ZSCP: 189 codon pairs not seen in the *C. reinhardtii* chloroplast genome

were predicted to have been present by the model. These were termed unexpectedly unseen zero scoring codon pairs (uuZSCPs).

- In **Hypothesis Seven,** the nine uuZSCPs with the highest predicted occurrences from the previous hypothesis were cross-referenced against expressing and non-expressing recombinant genes. Six of the uuZSCPs were found to be present in at least one expressing transgene, showing that these, too, are tolerated.
- **Hypothesis Eight** looked into the possibility that local regions of bad codon use, rather than point or gene wide effects, were related to the expression or non-expression of transgenes. Both optimised and non-optimised sets of recombinant genes were examined, but no comprehensive patterns were observed.
- **Hypothesis Nine** investigated whether ZSCPs are conserved between the chloroplasts of 14 related green algal species. Five codon pairs were found to be absent from all 14; however, the rarity of these codon pairs based on *C. reinhardtii* chloroplast codon preferences and amino acid pair usage negated any significance of this finding.
- Finally, **Hypothesis Ten** analysed the limited tRNA pool seen in the *C. reinhardtii* chloroplast in regard to relative codon preferences. Surprisingly, it was found that codon preference does not follow tRNA availability. In fact, there is a wide range of tRNA related preference seen, from 94 % preference for the native tRNA codon for lysine to 90 % preference for wobble factor codons for cysteine.

To conclude, it has been shown that codon pair preference does exists in the *C. reinhardtii* chloroplast genome; however, so far no data has been presented to link either codon-, or codon pair optimisation to absolute recombinant gene expression. It is likely that this finding is a result of the absolute nature of the expression investigated, as for protein accumulation to be entirely undetectable, it is likely that several detrimental factors are working together. An interesting future area of study would be to look at occurrence of rare codon pairs in relation to levels of expressed proteins. To take a more experimental approach, as opposed to the observational study presented above, a prime target for such an investigation is the uuZSCP UUU:CGU (Phe:Arg) in *cpl-1*. Expression of *cpl-1* is

considered high even with this codon pair, but a convincing study into whether such actively avoided codon pairs are detrimental would be to make a silent mutation at this site and observe the effect, if any, on Cpl-1 accumulation.

As commented on in Hypothesis seven, the datasets used to investigate expressing and non-expressing transgenes were very small. In order to conduct a more robust investigation into the effect of codon use on recombinant gene expression, a considerably larger sample would be required. This has thus far been hindered by the tendency for optimised gene sequences not to be published, and non-expressing genes not to be reported at all. A collaborative approach where a number of labs were contacted directly could greatly increase the number of transgenes available for investigation, and thus the significance of data collected.

It is also possible that, as has been suggested for other organisms, that 'good' codon- and codon pair use has a greater effect on the overall cell fitness than gene-specific expression. This could also be investigated experimentally by comparing growth rates of *C. reinhardtii* lines transformed with 'good' and 'bad' versions of the same gene. Care would have to be taken in such an experiment, however, to ensure that the genes selected were expressed to a similar level.

A final area of potential study which has become apparent during this project is that of unexpectedly favoured codon pairs: those which are seen considerably more often than is predicted by the non codon biased *p(dataset3)*. Understanding why such preferences occur may not be as directly relevant as that of the actively avoided codons, (as the effects on transgene expression is likely to be more subtle); however, as has become increasingly evident throughout this and the preceding chapter, the process of foreign gene expression in the *C. reinhardtii* chloroplast is incredibly multifaceted, and the more that can be understood about the basic biology involved, the greater the chance of success.

Chapter 6

General Discussion

6.1 The synthesis of the *Streptococcus pneumoniae* endolysin Cpl-1 in the chloroplast of *Chlamydomonas reinhardtii*

6.1.1 Research presented

Chapter three describes the design and successful expression in the algal chloroplast of a synthetic gene encoding the bacteriophage endolysin Cpl-1. The lysin was isolated following a protocol adapted from earlier work on the expression of *cpl-1* in *E. coli* (Loeffler *et al.*, 2001) and produced a greatly enriched Cpl-1 preparation. The activity of the recombinant lysin against its target, *S. pneumoniae,* was investigated in liquid culture by a turbidity clearance assay where bacterial cell lysis was measured by a drop in optical density at 600 nm. Both crude and enriched extracts displayed significant clearance of bacterial cultures relative to both buffer and non Cpl-1 containing extracts. Furthermore, the rate of clearance was shown to be significantly faster for the enriched extract, which contained a higher concentration of Cpl-1 as demonstrated by western blot analysis. *C. reinhardtii* synthesised Cpl-1 was shown to be active against three out of four clinical isolates of *S. pneumoniae;* however, no activity was observed against *S. pyogenes* or *E. coli*.

Preliminary investigations into Cpl-1 stability in a *C. reinhardtii* crude cell extract showed little degradation after 93 hours at 4 °C, a steadily decline in Cpl-1 concentration over this time period at 25 °C, and rapid degradation over 24 hours at 37 °C. Accumulation of Cpl-1 under the control of the *psaA* promoter/ 5' UTR was quantified by western blot analysis comparison with a commercial standard and a figure of approximately 9 % TSP calculated, although this requires verification.

6.1.2 Future prospects

Further investigations into the expression of *cpl-1* in the *C. reinhardtii* chloroplast can be split into two categories: the continuation of this project by further investigation into the potential of native Cpl-1 as an antimicrobial, and the application of synthetic biology to produce a therapeutic more tailored towards clinical use.

6.1.2.1 ***Continuing investigations into native Cpl-1***

In order to properly conclude this project there are several objectives that should be addressed, firstly the quantification of Cpl-1 accumulating in the *C. reinhardtii* chloroplast. This is important both for conducting feasibility studies in relation to projected costings of a commercial product, and also for further process optimisation where accurate quantification throughout the purification protocol will be necessary. It has been suggested that the discrepancy between western blot analysis and Coomassie stained comparisons between samples and standards could be related to the purified nature of the standard. A potential future strategy would thus be to analyse the standard in the presence of a blank *C. reinhardtii* sample. Alternatively quantification could be conducted by ELISA as opposed to western blot analysis.

Secondly, it is important to derive a strategy capable of producing a truly purified (as opposed to simply enriched) Cpl-1 sample. Such a sample could then be used to conduct enzymatic studies on Cpl-1 to ensure that the *C. reinhardtii* synthesised lysin shows a comparable specific activity relative to that produced by others in *E. coli* or plant based systems (Loeffler *et al.*, 2003; Oey *et al.*, 2009b). The current purification strategy, though involving a matrix typically used for ion exchange chromatography, could more actively be described as affinity chromatography as DEAE acts as a choline analogue. Given the extreme binding affinity of Gram-positive lysins for their cell wall target, this technique should thus be able to produce highly pure Cpl-1 extracts, once properly optimised. Optimisation could be conducted by first using a NaCl gradient to ascertain the concentration at which Cpl-1 elutes from the column, and then tailoring future wash steps accordingly.

Once purified Cpl-1 samples of known concentration are produced, enzymatic characterisation can be conducted to give values for specific activity as well as Michaelis-Menten kinetic data if a suitable substrate can be acquired. Such purified Cpl-1 would also be required for the progression of *in vivo* activity analysis in animal models, with a logical starting point being to replicate early Cpl-1 studies conducted on murine pneumococcal colonisation models (Loeffler *et al.*, 2003).

6.1.2.2 ***Application of synthetic biology to Cpl-1***

There is considerable interest in the modification of natural lysins to generate novel proteins with improved clinical relevance. Such developments include increasing catalytic activity, broadening pathogen target range, and reducing immunogenicity (Schmelcher *et al.*, 2012). One particular area of interest has been boosting catalytic activity by directly reducing the efficiency of the cell binding domain, or removing it altogether. As discussed above, many Gram-positive lysins have nanomolar range affinity constants for their substrates. This in essence means that each lysin molecule is single use: it binds to the cell wall, cleaves a single peptidoglycan bond, and then remains attached (Korndörfer *et al.*, 2006; Loessner *et al.*, 2002). This characteristic, while advantageous for phage progeny in a native context, is undesirable in a therapeutic. The discrete modular nature of many lysins can be highly amenable to modifications such as domain swapping and truncation, with several groups showing increases in activity over the natural enzyme from such alterations (Cheng and Fischetti, 2007; Horgan *et al.*, 2009; Schmelcher *et al.*, 2011). Cpl-1 activity however, has been shown to be severely compromised on removal of its cell-binding domain (CBD), giving rise to the suggestion that the CBD is required for positioning of the catalytic domain (CD) relative to the substrate (Sanz *et al.*, 1992). In order to reduce the cell wall affinity of Cpl-1 such that enzyme turnover is increased without having a detrimental effect on CD positioning, the choline binding activity of the CBD would have to be weakened, rather than removed or replaced completely. From the crystal structure of Cpl-1 it has been suggested that of the four predicted choline binding sites present on the CDB, only two are active (Hermoso *et al.*, 2003). The choline methyl groups are each stabilised by a conserved aromatic residue; replacement of one or more of these interactions may be sufficient to disrupt the site altogether, restricting the Cpl-1 CBD to a single choline-binding site. By theoretically reducing the cell wall binding specificity it is possible that pathogen target range would also be increased, as is observed for the PlyL lysin (Low *et al.*, 2005). This could also be advantageous as discussed below.

6.2 Attempts to express other lysins, and a study into the problems of foreign gene expression in the *C. reinhardtii* chloroplast

6.2.1 Research presented

In Chapter four, synthetic genes encoding two further lysins, (Gp20 and Lys16) were designed and successfully introduced into the *C. reinhardtii* chloroplast. However, neither transformant line gave detectable levels of recombinant protein accumulation, and thus the focus of the chapter was turned towards the improvement of transgene expression by translation optimisation. A novel expression cassette was used to generate $atpA_{34}$*:GoI* chimeras, faithfully recreating the translation initiation region from the endogenous *atpA* gene. The chimeric genes did not give detectable expression; however, work by others implied this might be due to instability of the gene product as opposed to a failure to improve translation initiation. In order to recreate a translation initiation region of proven function while maintaining folding stability, and to simultaneously create dual functioning lysins, full enzyme fusions were generated. The *gp20* and *lys16* sequences were fused to an upstream copy of *cpl-1* via a flexible linker region, with a *cpl-1:pal* fusion as a positive control. Of the three fusion genes, only the positive control, *cpl-1:pal* showed detectable expression in the *C. reinhardtii* chloroplast suggesting that translation was compromised by ribosome stalling within the *gp20* and *lys16* regions of the transcripts. The Cpl-1:Pal fusion lysin was purified by choline affinity chromatography, demonstrating that at least one cell-binding domain is correctly folded; however, no antimicrobial activity was observed. The novel optimisation software, CUO, was utilised to redesign *gp20* with improved codon- and codon pair adaptation, but this also gave no detectable expression.

6.2.2 Future prospects

6.2.2.1 *Redesign of the fusion lysin flexi-linker*

An advantage of the lysins over conventional antibiotics is that they do not disrupt the natural microflora due to their highly specific nature (Fischetti, 2003). Though beneficial from a clinical prospective, this property is considered disadvantageous from a pharmaceutical standpoint: an agent that targets a very narrow range of stains costs the same to license and produce as one that can treat an entire phylum

(David Harper, AmpliPhi Biosciences, personal communication). An agent that could target multiple pathogens while remaining benign towards commensal bacteria would thus be the ideal solution. The purification of the Cpl-1:Pal fusion lysin by selective binding to the choline analogue DEAE has shown that at least one of the two choline binding domains present is correctly folding. The lack of activity from either domain suggests issues involving steric hindrance as opposed to mis-folding. Cpl-1 for example, is known to dimerise via the CBD for full activity (Resch *et al.*, 2011b); such an event could conceivably be blocked by the presence of Pal. Two potential solutions to this problem involve the flexible linker connecting the two lysins, and are illustrated in Figure 6.1. The first is to extend the 14-mer flexi-linker region to allow the two enzymes to interact more independently. This could restore activity, however it is possible that the presence of a longer unstructured region may act to recruit proteases. An alternative to a longer linker is to actively promote cleavage of the two enzymes. This could be achieved by the addition of the stromal possessing peptidase region seen in the pASap2 vector (Appendix j). The two enzymes would be cleaved within the chloroplast producing a mix of Pal and Cpl-1 as opposed to a single protein containing both enzymes. The latter solution would be of particular interest for the Gp20 and Lys16 lysins where expression in the *C. reinhardtii* chloroplast has yet to be observed.

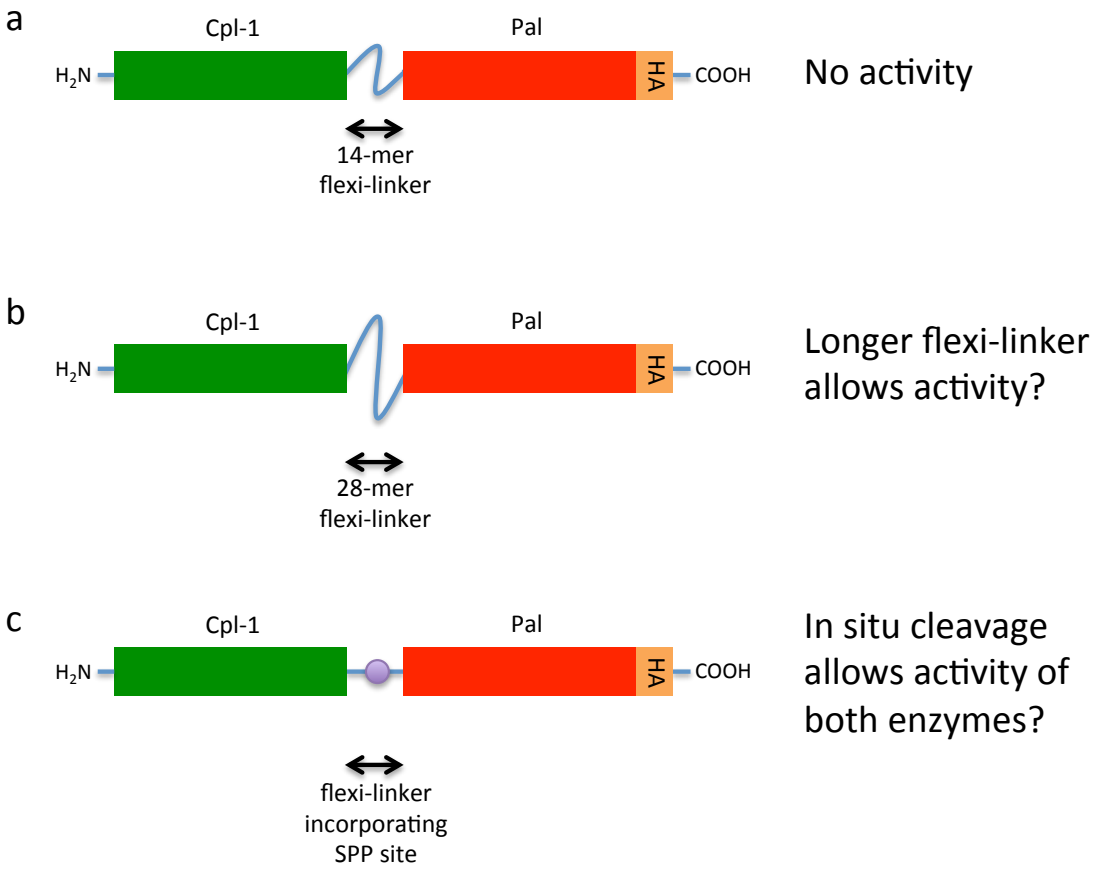


Figure 6.1 – Potential modifications to the Cpl-1:Pal fusion lysin to promote activity

a) The current Cpl-1:Pal fusion is expressed but fails to exhibit enzymatic activity, possible due to steric hindrance issues. **b)** The extension of the 14-mer flexi-linker between the two enzymes might allow for independent activity. **c)** The addition of a stromal processing peptidase (SPP) site would allow cleavage of the two enzymes producing a 'lysin cocktail' as opposed to a 'cocktail lysin'.

6.2.2.2 Further investigations into the expression of atpA$_{34}$:GoI *chimeras*

The detrimental effect of the N-terminal extension seen for AtpA34:Cpl-1 was discussed in Chapter four. It was suggested that the positioning of the tight alpha helical structure of the AtpA$_{34}$ moiety might destabilise the Cpl-1 protein due to the semi-buried position of the N-terminus between the two domains. There is currently no structural data available for Gp20 or Lys16 so it is not possible to comment on whether a similar disruption is occurring; however, there is a crystal structure for another protein previously expressed recombinantly in the *C. reinhardtii* chloroplast in the Purton lab. Human Growth Hormone (hGH), also known as somatotropin, is a small (22.3 kDa) globular protein that is used therapeutically to treat a number of growth hormone based deficiencies (Mehta and Hindmarsh, 2002). The crystal structure was solved in 1995 (Chantalat *et al.*, 1995) and shows the N-terminus to be exposed, and thus would make for an interesting candidate for further investigating the nature of the translation initiation region in *C. reinhardtii* chloroplast transgene expression (Figure 6.2).

It has been observed that the C-terminus of Cpl-1 is also located in the region between the two domains; however, extensions such as the HA tag and flexi-linker of the fusion lysin Cpl-1:Pal do not seem to have generated instability. It is likely that this is due to the nature of the extensions – the flexi-linker was specifically designed to form a random coil conformation, and the HA tag has be modelled and again shows little secondary structure (data not shown). If expression of an *atpA$_{34}$:hGH* gave improved accumulation of recombinant protein, a next step could be the addition of a flexi-linker between the SPP and the GoI as shown in Figure 6.3.

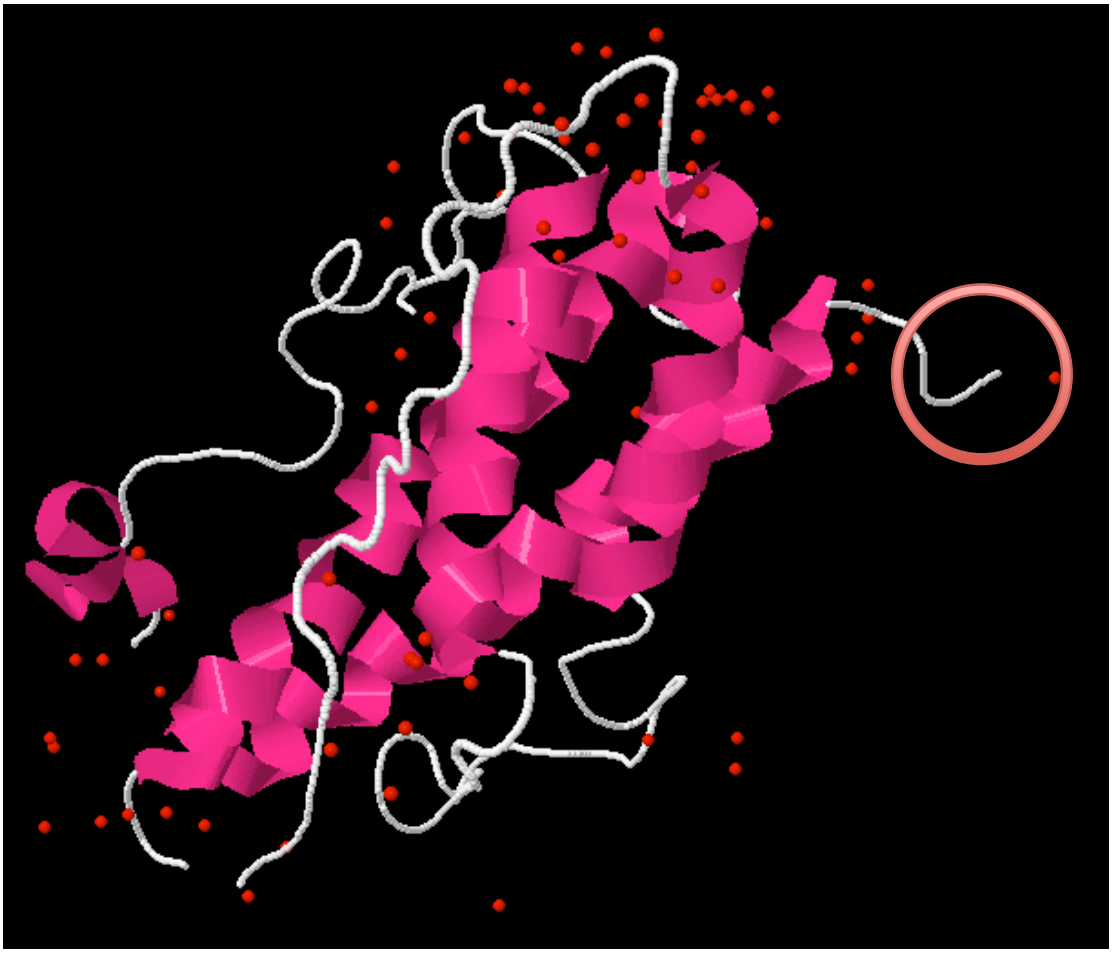


Figure 6.2 – The crystal structure of human Growth Hormone highlighting the N-terminus

Unlike Cpl-1 above, the N-terminus of human Growth Hormone is shown to be separate from the rest of the structure, making it theoretically more suitable for *atpA*$_{34}$ chimeric expression. Protein structure retrieved from the Protein Data Bank (PDB identifier 1HGU), and visualised using Jmol.

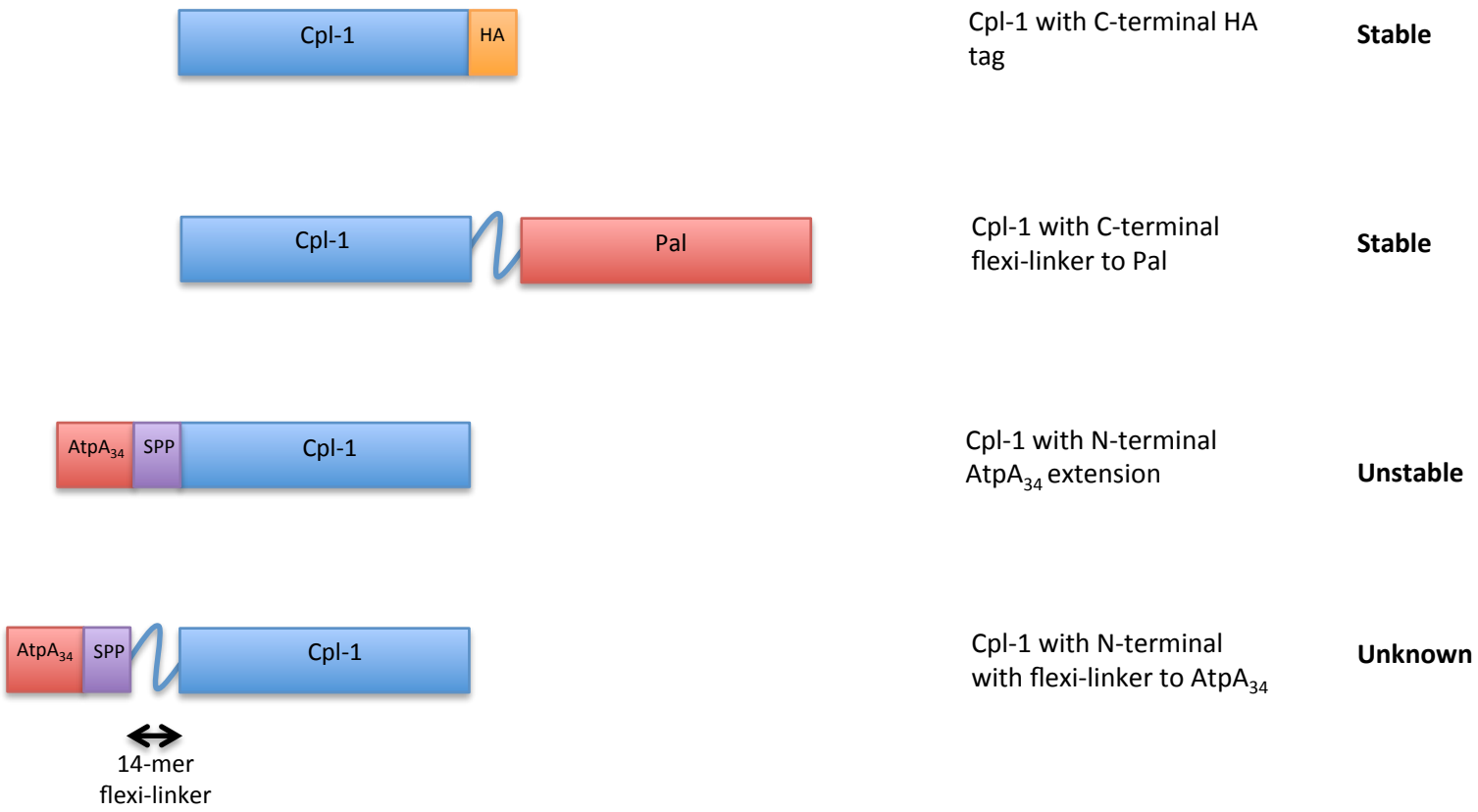


Figure 6.3 – The addition of a 14-mer flexi-linker to improve the stability of AtpA$_{34}$:PoI chimeras

The crystal structure data presented in Chapter four suggested that the instability of the AtpA$_{34}$:Cpl-1 chimera may be due to the positioning of the AtpA$_{34}$ moiety. The introduction of a flexible linker region between the SSP site and Cpl-1 may resolve this issue.

6.3 A bioinformatics investigation into codon- and codon pair use in the *C. reinhardtii* chloroplast

6.3.1 Research presented

Chapter five investigated the codon- and codon pair preferences of the *C. reinhardtii* chloroplast genome from a bioinformatics prospective, and related these findings back to transgenes introduced into the chloroplast in the Purton lab. It was found that, although definite codon- and codon pair biases do exist in this genome, these preferences do not appear to show any correlation with whether a transgene is expressed or not.

6.3.2 Future prospects

6.3.2.1 ***Continuation of current analyses***

Two apparent limitations of the codon investigation presented above are the size of the transgene pool investigated and the qualitative (as opposed to quantitative), nature of the analysis. The former is a consequence of the preliminary nature of the study; only transgenes from the Purton lab were included. Now that this early study has been concluded and a codon pair bias demonstrated it would seem appropriate to start contacting other labs in order to expand the transgene sample size. Introducing a quantitative dimension to the study may prove to be more challenging due to the issues with quantification discussed in Chapter three, especially when the involvement of other groups is taken into account. The best approach would likely be to group transgenes into discrete sets, the most rudimentary application being simply the discrimination between 'high' and 'low' levels of recombinant protein accumulation. From this starting point the model could be improved upon as more robust quantification techniques became available.

A further drawback to the research conducted in Chapter five is the absolute way in which rare codon pairs were considered, in that only pairs entirely unseen were investigated. As shown by Weiβ and colleagues, rare but not unseen codons can be sufficient to significantly disrupt the synthesis of the D1 protein (Weiß *et al.*, 2012). It would therefore be interesting to filter the data to identify codons that

are presence in the *C. reinhardtii* chloroplast genome, but not seen in transgenes of detectable expression.

6.3.2.2 *Site directed mutagenesis of the uuZSCP in* cpl-1

One of the most intriguing findings made during the course of this work was that the most strongly avoided unexpectedly unseen zero scoring codon pair (uuZSCP) is in fact present in the well expressed *cpl-1* gene. The conclusion drawn from this finding was that either uuZSCPs are avoided for reasons completely separate from translational elongation such as relating to overall cell fitness, or that Cpl-1 accumulates to high levels *despite* the presence of this potentially detrimental codon pair. Which of these is the case could be investigated experimentally by replacing the uuZSCP UUU:CGU with UUC:CGU and then re-introducing the gene into the *C. reinhardtii* chloroplast. The UUC:CGU codon pair encodes 32 of the 58 total occurrences of the Phe:Arg amino acid pair thus is clearly well tolerated. An increase in Cpl-1 accumulation as demonstrated by western blot analysis would be strong evidence that the UUU:CGU pair is detrimental to translation in the *C. reinhardtii* chloroplast.

6.3.2.3 An experimental approach to codon pair investigation utilizing the D1 loop region

In order to conduct a broader experimental investigation into how codon pair usage affects gene expression in the *C. reinhardtii* chloroplast a similar strategy to that demonstrated by Weiβ and colleagues could be employed (Weiß *et al.*, 2012). It has been shown that the flexible loop region of D1, encoded by *psbA*, can be altered without affecting the folding or function of the protein. This fact was utilised by Weiβ and colleagues to investigate rare codons, with detrimental effects being demonstrated by an inability to rescue a *ΔpsbA* mutant. Though effective for the purpose of the study, this approach only gives data on whether or not expression of *psbA* is sufficient to allow photosynthetic growth, and thus assumes failure to isolate transformants to be proof of low expression. A more thorough strategy might be to co-transform the *ΔpsbA* mutant with another selectable marker, for example *aadA* for spectinomycin/ streptomycin resistance. Once transformants are generated for each gene variant, accumulation of D1 could then

by assayed by western blot analysis using anti-D1 antibodies. Photosynthetic activity could also be analysed in order to give a comparison to the data presented by Weiβ and colleagues.

6.4 Concluding remarks

This study has demonstrated the first reported case of the algal chloroplast as an expression platform for the production of a viral protein antibiotic. As predicted from the native stability of lysins in prokaryotic-like environments (Fischetti *et al.*, 2006; Wang *et al.*, 2000) and performance in the chloroplast of higher plants (Oey *et al.*, 2009b), the Cpl-1 lysin accumulates to relatively high levels in the *C. reinhardtii* chloroplast, and has been shown to fold into an active conformation. The demonstration of successful production of a bioactive protein antibiotic in the potentially low cost expression platform *C. reinhardtii* marks a step forward in the adoption of the lysins as next generation antibiotics; a process which can be considered to be held back by the high cost associated with biologic therapeutics (Schmelcher *et al.*, 2012). The potential for the *C. reinhardtii* chloroplast as a platform for the continued development of synthetic lysins has also been demonstrated by the stable accumulation of the Cpl-1:Pal fusion lysin.

The work presented here has furthered our understanding of transgene expression in the *C. reinhardtii* chloroplast. The data presented in Chapter four suggests that, contrary to previous reports (Coragliotti *et al.*, 2011), translation initiation might not necessarily be the limiting factor in recombinant protein synthesis. This research also represents the first reported investigation into codon pair preference in the *C. reinhardtii* chloroplast genome, or indeed in any other organelle. Such a bias has been shown to exist, and further experiments proposed to ascertain the effects of unfavourable codon pairs on transgene expression.

It is likely that as conventional antibiotic reserves become increasingly depleted, and the use of protein based therapeutics more commonplace, lysins will become a part of our antimicrobial repertoire. The proof of concept presented here provides a compelling argument for the *C. reinhardtii* chloroplast as an expression platform for this new generation of therapeutics.

References

Abraham, E.P., Chain, E., 1940. An Enzyme from Bacteria able to Destroy Penicillin. Nature 146, 837–837.

Adam, Z., Rudella, A., van Wijk, K.J., 2006. Recent advances in the study of Clp, FtsH and other proteases located in chloroplasts. Curr. Opin. Plant Biol. 9, 234–240.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Antignac, A., Tomasz, A., 2009. Reconstruction of the phenotypes of methicillin-resistant Staphylococcus aureus by replacement of the staphylococcal cassette chromosome mec with a plasmid-borne copy of Staphylococcus sciuri pbpD gene. Antimicrob. Agents Chemother. 53, 435–441.

Barnes, D., Franklin, S., Schultz, J., Henry, R., Brown, E., Coragliotti, A., Mayfield, S., 2005. Contribution of 5′- and 3′-untranslated regions of plastid mRNAs to the expression of Chlamydomonas reinhardtii chloroplast genes. Mol. Genet. Genomics 274, 625–636.

Bateman, J.M., Purton, S., 2000. Tools for chloroplast transformation in Chlamydomonas: expression vectors and a new dominant selectable marker. Mol. Gen. Genet. 263, 404–410.

Bertani, G., 1951. Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. J. Bacteriol. 62, 293–300.

Blasi, F., Mantero, M., Santus, P., Tarsia, P., 2012. Understanding the burden of pneumococcal disease in adults. Clin. Microbiol. Infect. 18, 7–14.

Bock, R., 2001. Transgenic plastids in basic research and plant biotechnology. J. Mol. Biol. 312, 425–438.

Bolam, D.N., Ciruela, A., McQueen-Mason, S., Simpson, P., Williamson, M.P., Rixon, J.E., Boraston, A., Hazlewood, G.P., Gilbert, H.J., 1998. Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. Biochem. J. 331, 775–781.

Borysowski, J., Gorski, A., 2010. Fusion to cell-penetrating peptides will enable lytic enzymes to kill intracellular bacteria. Med. Hypotheses 74, 164–166.

Boycheva, S., Chkodrov, G., Ivanov, I., 2003. Codon pairs in the genome of Escherichia coli. Bioinformatics 19, 987–998.

Boynton, J.E., Gillham, N.W., Harris, E.H., Hosler, J.P., Johnson, A.M., Jones, A.R., Randolph-Anderson, B.L., Robertson, D., Klein, T.M., Shark, K.B., Sanford, J.C., 1988. Chloroplast transformation in chlamydomonas with high velocity microprojectiles. Science, New Series 240, 1534–1538.

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. 72, 248–254.

Brook, I., Frazier, E.H., 1991. Infections caused by propionibacterium species. Rev. Infect. Dis. 13, 819–822.

Buchanan, M.J., Imam, S.H., Eskue, W.A., Snell, W.J., 1989. Activation of the cell wall degrading protease, lysin, during sexual signalling in Chlamydomonas: the enzyme is stored as an inactive, higher relative molecular mass precursor in the periplasm. J. Cell Biol. 108, 199 –207.

Chantalat, L., Jones, N., Korber, F., Navaza, J., Pavlovsky, A., 1995. The crystal-structure of wild-type growth-hormone at 2.5 angstrom resolution. Protein Pept. Lett. 2, 333–340.

Cheng, Q., Fischetti, V.A., 2007. Mutagenesis of a bacteriophage lytic enzyme PlyGBS significantly increases its antibacterial activity against group B streptococci. Appl. Microbiol. Biotechnol. 74, 1284–1291.

Coates, T., Bax, R., Coates, A., 2009. Nasal decolonization of Staphylococcus aureus with mupirocin: strengths, weaknesses and future prospects. J Antimicrob Chemother 64, 9–15.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., Mueller, S., 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320, 1784–1787.

Cookson, B.D., 1998. The emergence of mupirocin resistance: a challenge to infection control and antibiotic prescribing practice. J. Antimicrob. Chemother. 41, 11–18.

Coragliotti, A., Beligni, M., Franklin, S., Mayfield, S., 2011. Molecular factors affecting the accumulation of recombinant proteins in the Chlamydomonas reinhardtii chloroplast. Mol. Biotechnol. 48, 60–75.

Corchero, J.L., Gasser, B., Resina, D., Smith, W., Parrilli, E., Vázquez, F., Abasolo, I., Giuliani, M., Jäntti, J., Ferrer, P., Saloheimo, M., Mattanovich, D., Schwartz Jr., S., Tutino, M.L., Villaverde, A., 2013. Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. Biotechnol. Adv. 31, 140–153.

Cullen, M., Ray, N., Husain, S., Nugent, J., Nield, J., Purton, S., 2007. A highly active histidine-tagged Chlamydomonas reinhardtii Photosystem II preparation for structural and biophysical analysis. Photochem. Photobiol. Sci. Off. J. Eur. Photochem. Assoc. Eur. Soc. Photobiol. 6, 1177–1183.

Davies, J., Davies, D., 2010. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 74, 417–433.

Díaz, E., López, R., García, J.L., 1990. Chimeric phage-bacterial enzymes: a clue to the modular evolution of genes. Proc. Natl. Acad. Sci. U. S. A. 87, 8125–8129.

Djurkovic, S., Loeffler, J.M., Fischetti, V.A., 2005. Synergistic killing of Streptococcus pneumoniae with the bacteriophage lytic enzyme Cpl-1 and penicillin or gentamicin depends on the level of penicillin resistance. Antimicrob Agents Chemother 49, 1225–1228.

Donskey, C.J., 2006. Antibiotic regimens and intestinal colonization with antibiotic-resistant Gram-negative bacilli. Clin. Infect. Dis. 43, S62–S69.

Dove, A., 2002. Uncorking the biomanufacturing bottleneck. Nat Biotech 20, 777–779.

Drawz, S.M., Bonomo, R.A., 2010. Three decades of beta-lactamase inhibitors. Clin. Microbiol. Rev. 23, 160–201.

Dreesen, I.A.J., Charpin-El Hamri, G., Fussenegger, M., 2010. Heat-stable oral alga-based vaccine protects mice from Staphylococcus aureus infection. J. Biotechnol. 145, 273–280.

Eady, E.A., Cove, J.H., Holland, K.T., Cunliffe, W.J., 1989. Erythromycin resistant propionibacteria in antibiotic treated acne patients - association with therapeutic failure. Br J Dermatol 121, 51–57.

Eaton, M.D., Bayne-Jones, S., 1934. Bacteriophage therapy. J. Am. Med. Assoc. 103, 1769 –1776.

Eberhard, S., Drapier, D., Wollman, F.-A., 2002. Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of Chlamydomonas reinhardtii. Plant J. 31, 149–160.

Entenza, J.M., Loeffler, J.M., Grandgirard, D., Fischetti, V.A., Moreillon, P., 2005. Therapeutic effects of bacteriophage Cpl-1 lysin against Streptococcus pneumoniae endocarditis in rats. Antimicrob Agents Chemother 49, 4789–4792.

Farrar, M.D., Howson, K.M., Bojar, R.A., West, D., Towler, J.C., Parry, J., Pelton, K., Holland, K.T., 2007. Genome sequence and analysis of a Propionibacterium acnes bacteriophage. J Bacteriol 189, 4161–4167.
Farrar, M.D., Ingham, E., 2004. Acne: Inflammation. Clin. Dermatol. 22, 380–384.
Finer, J.J., Finer, K.R., Ponappa, T., 1999. Particle bombardment mediated transformation. Curr. Top. Microbiol. Immunol. 240, 59–80.
Finnis, C.J., Payne, T., Hay, J., Dodsworth, N., Wilkinson, D., Morton, P., Saxton, M.J., Tooth, D.J., Evans, R.W., Goldenberg, H., Scheiber-Mojdehkar, B., Ternes, N., Sleep, D., 2010. High-level production of animal-free recombinant transferrin from Saccharomyces cerevisiae. Microb. Cell Factories 9, 87.
Fischer, R., Schillberg, S., Hellwig, S., Twyman, R.M., Drossard, J., 2012. GMP issues for recombinant plant-derived pharmaceutical proteins. Biotechnol. Adv. 30, 434–439.
Fischetti, V.A., 2001. Phage antibacterials make a comeback. Nat Biotech 19, 734–735.
Fischetti, V.A., 2003. Novel method to control pathogenic bacteria on human mucous membranes, in: Chiorazzi, N., Lahita, R.G., Capra, J.D., Ferrarini, M., Zabriskie, J.B. (Eds.), Immune Mechanisms and Disease. New York Acad Sciences, pp. 207–214.
Fischetti, V.A., 2008. Bacteriophage lysins as effective antibacterials. Curr. Opin. Microbiol. 11, 393–400.
Fischetti, V.A., 2010. Bacteriophage endolysins: A novel anti-infective to control Gram-positive pathogens. Int. J. Med. Microbiol. 300, 357–362.
Fischetti, V.A., Nelson, D., Schuch, R., 2006. Reinventing phage therapy: are the parts greater than the sum? Nat. Biotechnol. 24, 1508–1511.
Foster, K., Saranak, J., Patel, N., Zarilli, G., Okabe, M., Kline, T., Nakanishi, K., 1984. A rhodopsin is the functional photoreceptor for phototaxis in the unicellular eukaryote Chlamydomonas. Nature 311, 756–759.
Franklin, S., Ngo, B., Efuet, E., Mayfield, S.P., 2002. Development of a GFP reporter gene for Chlamydomonas reinhardtii chloroplast. Plant J. 30, 733–744.
Fuhrmann, M., Oertel, W., Hegemann, P., 1999. A synthetic gene coding for the green fluorescent protein (GFP) is a versatile reporter in Chlamydomonas reinhardtii†. Plant J. 19, 353–361.
Garcia, J.L., Garcia, E., Arraras, A., Garcia, P., Ronda, C., Lopez, R., 1987. Cloning, purification, and biochemical characterization of the pneumococcal bacteriophage Cp-1 lysin. J Virol 61, 2573–2580.
Goldschmidtclermont, M., 1991. Transgenic Expression of Aminoglycoside Adenine Transferase in the Chloroplast - a Selectable Marker for Site-Directed Transformation of Chlamydomonas. Nucleic Acids Res. 19, 4083–4089.
Grandgirard, D., Loeffler, J.M., Fischetti, V.A., Leib, S.L., 2008. Phage lytic enzyme Cpl-1 for antibacterial therapy in experimental pneumococcal meningitis. J Infect Dis 197, 1519–1522.
Gray, B.N., Yang, H., Ahner, B.A., Hanson, M.R., 2011. An efficient downstream box fusion allows high-level accumulation of active bacterial beta-glucosidase in tobacco chloroplasts. Plant Mol. Biol. 76, 345–355.
Gray, M., Doolittle, W., 1982. Has the Endosymbiont Hypothesis Been Proven. Microbiol. Rev. 46, 1–42.
Gregory, J.A., Topol, A.B., Doerner, D.Z., Mayfield, S., 2013. Alga-produced cholera toxin-Pfs25 fusion proteins as oral vaccines. Appl. Environ. Microbiol. 79, 3917–3925.

Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., Welch, M., 2012. Engineering genes for predictable protein expression. Protein Expr. Purif. 83, 37–46.
Gutman, G.A., Hatfield, G.W., 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc. Natl. Acad. Sci. 86, 3699–3703.
Hanahan, D., 1983. Studies on transformation of Escherichia coli with plasmids. J. Mol. Biol. 166, 557–580.
Harris, E.H., 1989. The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and Its Laboratory Use. Academic Press.
Harris, E.H., Boynton, J.E., Gillham, N.W., 1994. Chloroplast ribosomes and protein synthesis. Microbiol. Rev. 58, 700–754.
He, D.-M., Qian, K.-X., Shen, G.-F., Zhang, Z.-F., Li, Y.-N., Su, Z.-L., Shao, H.-B., 2007. Recombination and expression of classical swine fever virus (CSFV) structural protein E2 gene in Chlamydomonas reinhardtii chroloplasts. Colloids Surfaces B-Biointerfaces 55, 26–30.
Heitzer, M., Eckert, A., Fuhrmann, M., Griesbeck, C., 2007. Influence of codon bias on the expression of foreign genes in microalgae. Transgenic Microalgae Green Cell Factories 616, 46–53.
Hermoso, J.A., Monterroso, B., Albert, A., Galán, B., Ahrazem, O., García, P., Martínez-Ripoll, M., García, J.L., Menéndez, M., 2003. Structural basis for selective recognition of pneumococcal cell wall by modular endolysin from phage Cp-1. Structure 11, 1239–1249.
Horgan, M., O'Flynn, G., Garry, J., Cooney, J., Coffey, A., Fitzgerald, G.F., Ross, R.P., McAuliffe, O., 2009. Phage lysin LysK can be truncated to its CHAP domain and retain lytic activity against live antibiotic-resistant staphylococci. Appl. Environ. Microbiol. 75, 872–874.
Hosler, J.P., Wurtz, E.A., Harris, E.H., Gillham, N.W., Boynton, J.E., 1989. Relationship between gene dosage and gene expression in the chloroplast of Chlamydomonas reinhardtii. Plant Physiol. 91, 648–655.
Housby, J.N., Mann, N.H., 2009. Phage therapy. Drug Discov. Today 14, 536–540.
Huang, C.-J., Lin, H., Yang, X., 2012. Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. J. Ind. Microbiol. Biotechnol. 39, 383–399.
Hudson, I.R.B., 1994. The efficacy of intranasal mupirocin in the prevention of staphylococcal infections: a review of recent experience. J. Hosp. Infect. 27, 81–98.
Huppert, M., MacPherson, D.A., Cazin, J., 1953. Pathogenesis of Candida albicans infection following antibiotic therapy. J. Bacteriol. 65, 171–176.
Hyams, J., Davies, D., 1972. Induction and Characterization of Cell-Wall Mutants of Chlamydomonas-Reinhardi. Mutat. Res. 14, 381–&.
Itakura, K., Hirose, T., Crea, R., Riggs, A., Heyneker, H., Bolivar, F., Boyer, H., 1977. Expression in Escherichia coli of a chemically synthesized gene for hormone Somatostatin. Science 198, 1056–1063.
Jado, I., López, R., García, E., Fenoll, A., Casal, J., García, P., 2003. Phage lytic enzymes as therapy for antibiotic-resistant Streptococcus pneumoniae infection in a murine sepsis model. J. Antimicrob. Chemother. 52, 967–973.
Kasai, S., Yoshimura, S., Ishikura, K., Takaoka, Y., Kobayashi, K., Kato, K., Shinmyo, A., 2003. Effect of coding regions on chloroplast gene expression in Chlamydomonas reinhardtii. J. Biosci. Bioeng. 95, 276–282.
Keeling, P.J., 2004. Diversity and evolutionary history of plastids and their hosts. Am. J. Bot. 91, 1481–1493.

Kindle, K.L., Richards, K.L., Stern, D.B., 1991. Engineering the chloroplast genome: techniques and capabilities for chloroplast transformation in Chlamydomonas reinhardtii. Proc. Natl. Acad. Sci. 88, 1721 –1725.

Kindle, K.L., Schnell, R.A., Fernández, E., Lefebvre, P.A., 1989. Stable nuclear transformation of Chlamydomonas using the Chlamydomonas gene for nitrate reductase. J. Cell Biol. 109, 2589 –2601.

Kinoshita, T., Fukuzawa, H., Shimada, T., Saito, T., Matsuda, Y., 1992. Primary structure and expression of a gamete lytic enzyme in Chlamydomonas reinhardtii: similarity of functional domains to matrix metalloproteases. Proc. Natl. Acad. Sci. 89, 4693 –4697.

Klinkert, B., Elles, I., Nickelsen, J., 2006. Translation of chloroplast psbD mRNA in Chlamydomonas is controlled by a secondary RNA structure blocking the AUG start codon. Nucleic Acids Res. 34, 386–394.

Kluytmans, J., van Belkum, A., Verbrugh, H., 1997. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. Clin Microbiol Rev 10, 505–520.

Korndörfer, I.P., Danzer, J., Schmelcher, M., Zimmer, M., Skerra, A., Loessner, M.J., 2006. The crystal structure of the bacteriophage PSA endolysin reveals a unique fold responsible for specific recognition of Listeria cell walls. J. Mol. Biol. 364, 678–689.

Kudla, G., Murray, A.W., Tollervey, D., Plotkin, J.B., 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324, 255–258.

Kutateladze, M., Adamia, R., 2010. Bacteriophages as potential new therapeutics to replace or supplement antibiotics. Trends Biotechnol. 28, 591–595.

Labrie, S.J., Samson, J.E., Moineau, S., 2010. Bacteriophage resistance mechanisms. Nat. Rev. Microbiol. 8, 317–327.

Leclercq, R., 2009. Epidemiological and resistance issues in multidrug-resistant staphylococci and enterococci. Clin Microbiol Infect 15, 224–231.

Li, Z., Kessler, W., van den Heuvel, J., Rinas, U., 2011. Simple defined autoinduction medium for high-level recombinant protein production using T7-based Escherichia coli expression systems. Appl. Microbiol. Biotechnol. 91, 1203–1213.

Llarrull, L.I., Testero, S.A., Fisher, J.F., Mobashery, S., 2010. The future of the [beta]-lactams. Curr. Opin. Microbiol. 13, 551–557.

Loeffler, J.M., Djurkovic, S., Fischetti, V.A., 2003. Phage lytic enzyme Cpl-1 as a novel antimicrobial for pneumococcal bacteremia. Infect Immun 71, 6199–6204.

Loeffler, J.M., Fischetti, V.A., 2003. Synergistic lethal effect of a combination of phage lytic enzymes with different activities on penicillin-sensitive and -resistant Streptococcus pneumoniae strains. Antimicrob Agents Chemother 47, 375–377.

Loeffler, J.M., Nelson, D., Fischetti, V.A., 2001. Rapid killing of Streptococcus pneumoniae with a bacteriophage cell wall hydrolase. Science 294, 2170–2172.

Loessner, M.J., Kramer, K., Ebel, F., Scherer, S., 2002. C-terminal domains of Listeria monocytogenes bacteriophage murein hydrolases determine specific recognition and high-affinity binding to bacterial cell wall carbohydrates. Mol. Microbiol. 44, 335–349.

Lohr, M., Schwender, J., Polle, J.E.W., 2012. Isoprenoid biosynthesis in eukaryotic phototrophs: A spotlight on algae. Plant Sci. 185, 9–22.

Lood, R., Collin, M., 2011. Characterization and genome sequencing of two Propionibacterium acnes phages displaying pseudolysogeny. BMC Genomics 12, 198.
Low, L.Y., Yang, C., Perego, M., Osterman, A., Liddington, R.C., 2005. Structure and lytic activity of a Bacillus anthracis prophage endolysin. J. Biol. Chem. 280, 35433–35439.
Lowy, F.D., 2003. Antimicrobial resistance: the example of Staphylococcus aureus. J. Clin. Invest. 111, 1265–1273.
Luzhetskyy, A., Pelzer, S., Bechthold, A., 2007. The future of natural products as a source of new antibiotics. Curr. Opin. Investig. Drugs 8, 608–613.
Ma, J.K.-C., Barros, E., Bock, R., Christou, P., Dale, P.J., Dix, P.J., Fischer, R., Irwin, J., Mahoney, R., Pezzotti, M., Schillberg, S., Sparrow, P., Stoger, E., Twyman, R.M., 2005. Molecular farming for new drugs and vaccines. EMBO Rep. 6, 593–599.
Manoharadas, S., Witte, A., Bläsi, U., 2009. Antimicrobial activity of a chimeric enzybiotic towards Staphylococcus aureus. J. Biotechnol. 139, 118–123.
Manuell, A.L., Beligni, M.V., Elder, J.H., Siefker, D.T., Tran, M., Weber, A., McDonald, T.L., Mayfield, S.P., 2007. Robust expression of a bioactive mammalian protein in Chlamydomonas chloroplast. Plant Biotechnol. J. 5, 402–412.
Marechal-Drouard, L., Weil, J.H., Dietrich, A., 1993. Transfer RNAs and Transfer RNA Genes in Plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 44, 13–32.
Marín-Navarro, J., Manuell, A., Wu, J., P. Mayfield, S., 2007. Chloroplast translation regulation. Photosynth. Res. 94, 359–374.
Marinelli, L.J., Fitz-Gibbon, S., Hayes, C., Bowman, C., Inkeles, M., Loncaric, A., Russell, D.A., Jacobs-Sera, D., Cokus, S., Pellegrini, M., Kim, J., Miller, J.F., Hatfull, G.F., Modlin, R.L., 2012. Propionibacterium acnes bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. mBio 3.
Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H., Stern, D.B., 2002. The Chlamydomonas reinhardtii plastid chromosome: Islands of genes in a sea of repeats. Plant Cell Online 14, 2659 –2679.
Mayer, M.J., Narbad, A., Gasson, M.J., 2008. Molecular characterization of a Clostridium difficile bacteriophage and its cloned biologically active endolysin. J. Bacteriol. 190, 6734–6740.
Mayfield, S.P., Franklin, S.E., Lerner, R.A., 2003. Expression and assembly of a fully active antibody in algae. Proc. Natl. Acad. Sci. 100, 438–442.
Mayfield, S.P., Manuell, A.L., Chen, S., Wu, J., Tran, M., Siefker, D., Muto, M., Marin-Navarro, J., 2007. Chlamydomonas reinhardtii chloroplasts as protein factories. Curr. Opin. Biotechnol. 18, 126–133.
Mayfield, S.P., Schultz, J., 2004. Development of a luciferase reporter gene, luxCt, for Chlamydomonas reinhardtii chloroplast. Plant J. 37, 449–458.
Mehta, A., Hindmarsh, P.C., 2002. The use of somatropin (recombinant growth hormone) in children of short stature. Paediatr. Drugs 4, 37–47.
Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., et al., 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318, 245 –250.
Monterroso, B., Sáiz, J.L., García, P., García, J.L., Menéndez, M., 2008. Insights into the structure-function relationships of pneumococcal cell wall lysozymes, LytC and Cpl-1. J. Biol. Chem. 283, 28618 –28628.
Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P., Bamberg, E., 2003. Channelrhodopsin-2, a directly light-gated

cation-selective membrane channel. Proc. Natl. Acad. Sci. U. S. A. 100, 13940–13945.
Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292.
Nelson, D., Loomis, L., Fischetti, V.A., 2001. Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. Proc. Natl. Acad. Sci. 98, 4107 –4112.
Nikaido, H., 1998. Antibiotic resistance caused by Gram-negative multidrug efflux pumps. Clin. Infect. Dis. 27, S32–S41.
O'Brien, K.L., Bronsdon, M.A., Dagan, R., Yagupsky, P., Janco, J., Elliott, J., Whitney, C.G., Yang, Y.-H., Robinson, L.-G.E., Schwartz, B., Carlone, G.M., 2001. Evaluation of a medium (STGG) for transport and optimal recovery of Streptococcus pneumoniae from nasopharyngeal secretions collected during field studies. J Clin Microbiol 39, 1021–1024.
O'Brien, K.L., Wolfson, L.J., Watt, J.P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O.S., Cherian, T., 2009. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. The Lancet 374, 893–902.
O'Connor, H.E., Ruffle, S.V., Cain, A.J., Deak, Z., Vass, I., Nugent, J.H.A., Purton, S., 1998. The 9-kDa phosphoprotein of photosystem II. Generation and characterisation of Chlamydomonas mutants lacking PSII-H and a site-directed mutant lacking the phosphorylation site. Biochim. Biophys. Acta-Bioenerg. 1364, 63–72.
O'Flaherty, S., Coffey, A., Meaney, W., Fitzgerald, G.F., Ross, R.P., 2005. The recombinant phage lysin LysK has a broad spectrum of lytic activity against clinically relevant Staphylococci, including methicillin-resistant Staphylococcus aureus. J. Bacteriol. 187, 7161–7164.
O'Flaherty, S., Ross, R.P., Coffey, A., 2009. Bacteriophage and their lysins for elimination of infectious bacteria. FEMS Microbiol. Rev. 33, 801–819.
Oey, M., Lohse, M., Kreikemeyer, B., Bock, R., 2009a. Exhaustion of the chloroplast protein synthesis capacity by massive expression of a highly stable protein antibiotic. Plant J 57, 436–445.
Oey, M., Lohse, M., Scharff, L.B., Kreikemeyer, B., Bock, R., 2009b. Plastid production of protein antibiotics against pneumonia via a new strategy for high-level expression of antimicrobial proteins. Proc. Natl. Acad. Sci. 106, 6579–6584.
Passwater, R., Solomon, N., 1997. Algae: The next generation of superfoods. Experts Optim. Heal. J. 1, 2–10.
Payne, D.J., Gwynn, M.N., Holmes, D.J., Pompliano, D.L., 2007. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat. Rev. Drug Discov. 6, 29–40.
Perl, T.M., Cullen, J.J., Wenzel, R.P., Zimmerman, M.B., Pfaller, M.A., Sheppard, D., Twombley, J., French, P.P., Herwaldt, L.A., 2002. Intranasal mupirocin to prevent postoperative staphylococcus aureus infections. N. Engl. J. Med. 346, 1871–1877.
Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42.
Potvin, G., Zhang, Z., 2010. Strategies for high-level recombinant protein expression in transgenic microalgae: A review. Biotechnol. Adv. 28, 910–918.

Purton, S., 2007. Tools and techniques for chloroplast transformation of Chlamydomonas, in: Transgenic Microalgae as Green Cell Factories, Advances in Experimental Medicine and Biology. Springer, pp. 34–45.

Purton, S., Szaub, J.B., Wannathong, T., Young, R., Economou, C.K., 2013. Genetic engineering of algal chloroplasts: Progress and prospects. Russ. J. Plant Physiol. 60, 491–499.

Ralston, D., Baer, B., Lieberman, M., Krueger, A., 1955. Virolysin - A virus-induced lysin from staphylococcal phage lysates. Proc. Soc. Exp. Biol. Med. 89, 502–507.

Rasala, B.A., Muto, M., Lee, P.A., Jager, M., Cardoso, R.M.F., Behnke, C.A., Kirk, P., Hokanson, C.A., Crea, R., Mendez, M., Mayfield, S.P., 2010. Production of therapeutic proteins in algae, analysis of expression of seven human proteins in the chloroplast of Chlamydomonas reinhardtii. Plant Biotechnol. J. 8, 719–733.

Rasala, B.A., Muto, M., Sullivan, J., Mayfield, S.P., 2011. Improved heterologous protein expression in the chloroplast of Chlamydomonas reinhardtii through promoter and 5' untranslated region optimization. Plant Biotechnol. J. 9, 674–683.

Rashel, M., Uchiyama, J., Ujihara, T., Uehara, Y., Kuramoto, S., Sugihara, S., Yagyu, K.-I., Muraoka, A., Sugai, M., Hiramatsu, K., Honke, K., Matsuzaki, S., 2007. Efficient elimination of multidrug-resistant Staphylococcus aureus by cloned lysin derived from bacteriophage ϕMR11. J. Infect. Dis. 196, 1237 – 1247.

Resch, G., Moreillon, P., Fischetti, V.A., 2011a. PEGylating a bacteriophage endolysin inhibits its bactericidal activity. AMB Express 1.

Resch, G., Moreillon, P., Fischetti, V.A., 2011b. A stable phage lysin (Cpl-1) dimer with increased antipneumococcal activity and decreased plasma clearance. Int. J. Antimicrob. Agents 38, 516–521.

Ringo, D., 1967. Flagellar Motion and Fine Structure of Flagellar Apparatus in Chlamydomonas. J. Cell Biol. 33, 543–&.

Roberts, K., Hills, G., Gurneysm.m, 1972. Structure, composition, and morphogenesis of the cell wall of Chlamydomonas reinhardi: Ultrastructure and preliminary chemical analysis. J. Ultrastruct. Res. 40, 599–&.

Rochaix, J.-D., 2001. Assembly, function, and dynamics of the photosynthetic machinery in Chlamydomonas reinhardtii. Plant Physiol. 127, 1394–1398.

Rochaix, J.-D., 2011. Regulation of photosynthetic electron transport. Biochim. Biophys. Acta-Bioenerg. 1807, 375–383.

Rochaix, J.D., Mayfield, S.P., Goldschmidt-Clermont, M., Erickson, J., 1988. Plant Molecular Biology. Oxford: IRL Press Limited.

Ronda, C., Lopez, R., Garcia, E., 1981. Isolation and characterization of a new bacteriophage, Cp-1, infecting Streptococcus pneumoniae. J Virol 40, 551–559.

Roque, A.C.A., Lowe, C.R., Taipa, M.A., 2004. Antibodies and genetically engineered related molecules: production and purification. Biotechnol. Prog. 20, 639–654.

Rosales-Mendoza, S., Teresita Paz-Maldonado, L.M., Elena Soria-Guerra, R., 2012. Chlamydomonas reinhardtii as a viable platform for the production of recombinant proteins: current status and perspectives. Plant Cell Reports 31, 479–494.

Rupprecht, J., 2009. From systems biology to fuel - Chlamydomonas reinhardtii as a model for a systems biology approach to improve biohydrogen production. J. Biotechnol. 142, 10–20.
Sanford, J.C., Smith, F.D., Russell, J.A., 1993. Optimizing the biolistic process for different biological applications. Methods Enzymol. 217, 483–509.
Sanz, J.M., Díaz, E., García, J.L., 1992. Studies on the structure and function of the N-terminal domain of the pneumococcal murein hydrolases. Mol. Microbiol. 6, 921–931.
Sass, P., Bierbaum, G., 2007. Lytic Activity of Recombinant Bacteriophage φ11 and φ12 Endolysins on Whole Cells and Biofilms of Staphylococcus aureus. Appl. Environ. Microbiol. 73, 347–352.
Schmelcher, M., Donovan, D.M., Loessner, M.J., 2012. Bacteriophage endolysins as novel antimicrobials. Future Microbiol. 7, 1147–1171.
Schmelcher, M., Tchang, V.S., Loessner, M.J., 2011. Domain shuffling and module engineering of Listeria phage endolysins for enhanced lytic activity and binding affinity. Microb. Biotechnol. 4, 651–662.
Schuch, R., Pelzek, A.J., Raz, A., Euler, C.W., Ryan, P.A., Winer, B.Y., Farnsworth, A., Bhaskaran, S.S., Stebbins, C.E., Xu, Y., Clifford, A., Bearss, D.J., Vankayalapati, H., Goldberg, A.R., Fischetti, V.A., 2013. Use of a bacteriophage lysin to identify a novel target for antimicrobial development. PLoS ONE 8, e60754.
Sharp, P.M., Li, W.H., 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.
Sheehan, M.M., Garcia, J.L., Lopez, R., Garcia, P., 1997. The lytic enzyme of the pneumococcal phage Dp-1: a chimeric lysin of intergeneric origin. Mol. Microbiol. 25, 717–725.
Sijmons, P.C., Dekker, B.M., Schrammeijer, B., Verwoerd, T.C., van den Elzen, P.J., Hoekema, A., 1990. Production of correctly processed human serum albumin in transgenic plants. Biotechnol. Nat. Publ. Co. 8, 217–221.
Smith, D., Yarus, M., 1989. tRNA-tRNA interactions within cellular ribosomes. Proc. Natl. Acad. Sci. 86, 4397–4401.
Snyder, J.W., Atlas, R.M., 2006. Handbook of Media for Clinical Microbiology. CRC Press.
Sonstein, S.A., Hammel, J.M., Bondi, A., 1971. Staphylococcal bacteriophage-associated lysin: a lytic agent active against Staphylococcus aureus. J Bacteriol 107, 499–504.
Specht, E., Miyake-Stoner, S., Mayfield, S., 2010. Micro-algae come of age as a platform for recombinant protein production. Biotechnol. Lett. 32, 1373–1383.
Specht, E.A., Mayfield, S.P., 2013. Synthetic oligonucleotide libraries reveal novel regulatory elements in Chlamydomonas chloroplast mRNAs. ACS Synth. Biol. 2, 34–46.
Spreitzer, R.J., Mets, L., 1981. Photosynthesis-deficient mutants of Chlamydomonas reinhardii with associated light-sensitive phenotypes. Plant Physiol. 67, 565–569.
Sprengart, M.L., Fuchs, E., Porter, A.G., 1996. The downstream box: an efficient and independent translation initiation signal in Escherichia coli. EMBO J. 15, 665–674.
Stern, D., Harris, E.H., 2009. The chlamydomonas sourcebook. Volume 2, Organellar and metabolic processes. Elsevier/Academic Press, Oxford, UK.

Sun, M., Qian, K., Su, N., Chang, H., Liu, J., Shen, G., 2003. Foot-and-mouth disease virus VP1 protein fused with cholera toxin B subunit expressed in Chlamydomonas reinhardtii chloroplast. Biotechnol. Lett. 25, 1087–1092.

Surzycki, Greenham, K., Kitayama, K., Dibal, F., Wagner, R., Rochaix, J.D., Ajam, T., Surzycki, S., 2009. Factors effecting expression of vaccines in microalgae. Biologicals 37, 133–138.

Suzuki, L., Johnson, C.H., 2001. Algae know the time of day: Circadian and photoperiodic programs. J. Phycol. 37, 933–942.

Swiech, K., Picanço-Castro, V., Covas, D.T., 2012. Human cells: A new platform for recombinant therapeutic protein production. Protein Expr. Purif. 84, 147–153.

Tran, M., Van, C., Barrera, D.J., Pettersson, P.L., Peinado, C.D., Bui, J., Mayfield, S.P., 2013. Production of unique immunotoxin cancer therapeutics in algal chloroplasts. Proc. Natl. Acad. Sci. 110, E15–E22.

Von Eiff, C., Becker, K., Machka, K., Stammer, H., Peters, G., 2001. Nasal carriage as a source of Staphylococcus aureus bacteremia. Study Group. N. Engl. J. Med. 344, 11–16.

Walker, T.L., Purton, S., Becker, D.K., Collet, C., 2005. Microalgae as bioreactors. Plant Cell Reports 24, 629–641.

Wang, I.-N., Smith, D.L., Young, R., 2000. Holins: The protein clocks of bacteriophage infections. Annu. Rev. Microbiol. 54, 799–825.

Wang, T., Li, Y., Shi, Y., Reboud, X., Darmency, H., Gressel, J., 2004. Low frequency transmission of a plastid-encoded trait in Setaria italica. Theor. Appl. Genet. 108, 315–320.

Wang, X., Brandsma, M., Tremblay, R., Maxwell, D., Jevnikar, A.M., Huner, N., Ma, S., 2008. A novel expression platform for the production of diabetes-associated autoantigen human glutamic acid decarboxylase (hGAD65). Bmc Biotechnol. 8.

Weiß, C., Bertalan, I., Johanningmeier, U., 2012. Effects of rare codon clusters on the expression of a high-turnover chloroplast protein in Chlamydomonas reinhardtii. J. Biotechnol. 160, 105–111.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., Gustafsson, C., 2009a. Design parameters to control synthetic gene expression in Escherichia coli. PLoS ONE 4, e7002.

Welch, M., Villalobos, A., Gustafsson, C., Minshull, J., 2009b. You're one in a googol: optimizing genes for protein expression. J. R. Soc. Interface R. Soc. 6 Suppl 4, S467–476.

Witzenrath, M., Schmeck, B., Doehn, J.M., Tschernig, T., Zahlten, J., Loeffler, J.M., Zemlin, M., Müller, H., Gutbier, B., Schütte, H., Hippenstiel, S., Fischetti, V.A., Suttorp, N., Rosseau, S., 2009. Systemic use of the endolysin Cpl-1 rescues mice with fatal pneumococcal pneumonia. Crit. Care Med. 37, 642–649.

Wu-Scharf, D., Jeong, B., Zhang, C., Cerutti, H., 2000. Transgene and transposon silencing in Chlamydomonas reinhardtii by a DEAH-box RNA helicase. Science 290, 1159–1162.

Yang, Z., Li, Y., Chen, F., Li, D., Zhang, Z., Liu, Y., Zheng, D., Wang, Y., Shen, G., 2006. Expression of human soluble TRAIL in Chlamydomonas reinhardtii chloroplast. Chin. Sci. Bull. 51, 1703–1709.

Yohn, C.B., Cohen, A., Rosch, C., Kuchka, M.R., Mayfield, S.P., 1998. Translation of the chloroplast psbA mRNA requires the nuclear-encoded poly(A)-binding protein, RB47. J. Cell Biol. 142, 435–442.

Ziedaite, G., Daugelavicius, R., Bamford, J.K.H., Bamford, D.H., 2005. The holin protein of bacteriophage PRD1 forms a pore for small-molecule and endolysin translocation. J Bacteriol 187, 5397–5405.

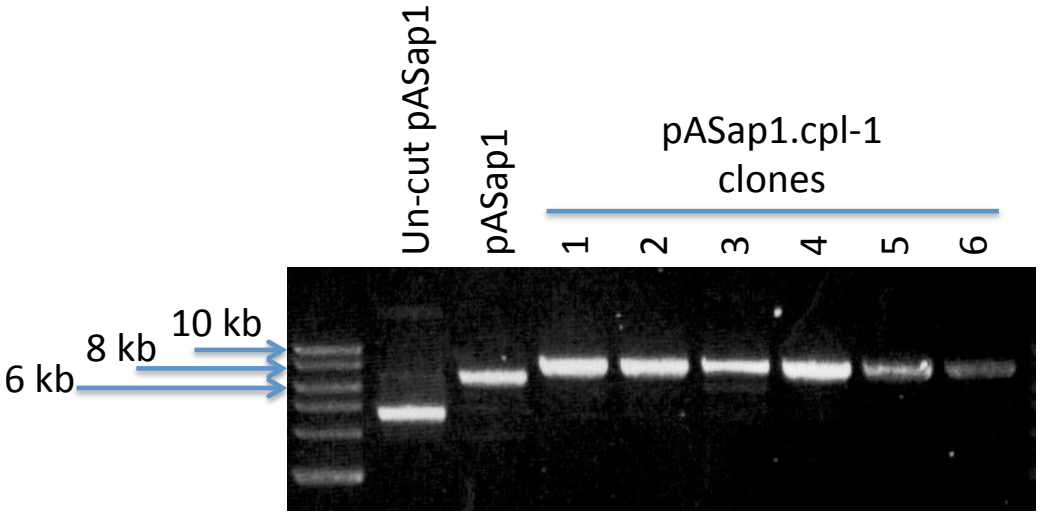
Zimmer, M., Sattelberger, E., Inman, R.B., Calendar, R., Loessner, M.J., 2003. Genome and proteome of Listeria monocytogenes phage PSA: an unusual case for programmed + 1 translational frameshifting in structural protein synthesis. Mol. Microbiol. 50, 303–317.

Appendices

Chapter three

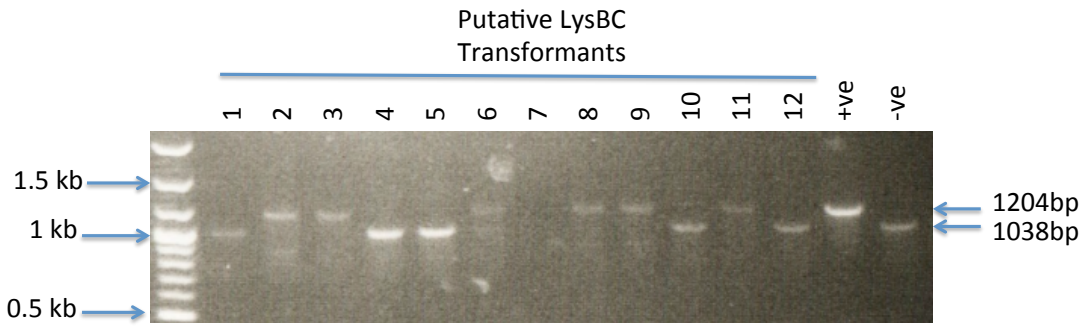
Appendix a - Confirmation of insertion of *cpl-1* into pASap1 by *Sph*I single digest

All six clones show the addition of approximately 1kb as expected for correct insertion of *cpl-1.*



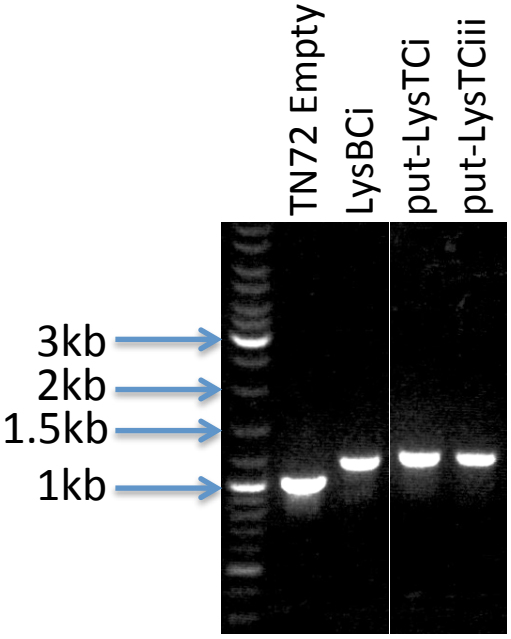
Appendix b - PCR screening of putative bst-same1 transformants for correct insertion of *cpl-1*

Of the 12 cell lines screened 6 show the correct band size of 1204bp indicating insertion of the GoI, whereas 6 show the wild type genotype indicating *psbH*-only transformation.



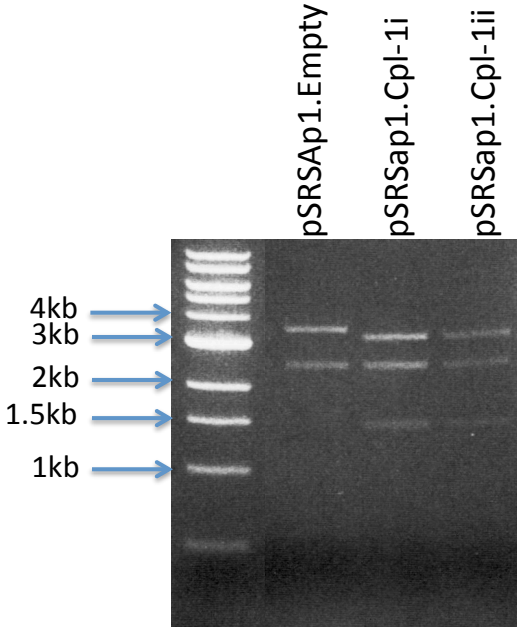
Appendix c– PCR screening of putative TN72 transformants for correct insertion of *cpl-1 (atpA* promoter)

Two transformant colonies were re-streaked on HSM and screened for the correct insertion of the expression cassette. Both were confirmed as correct transformants.



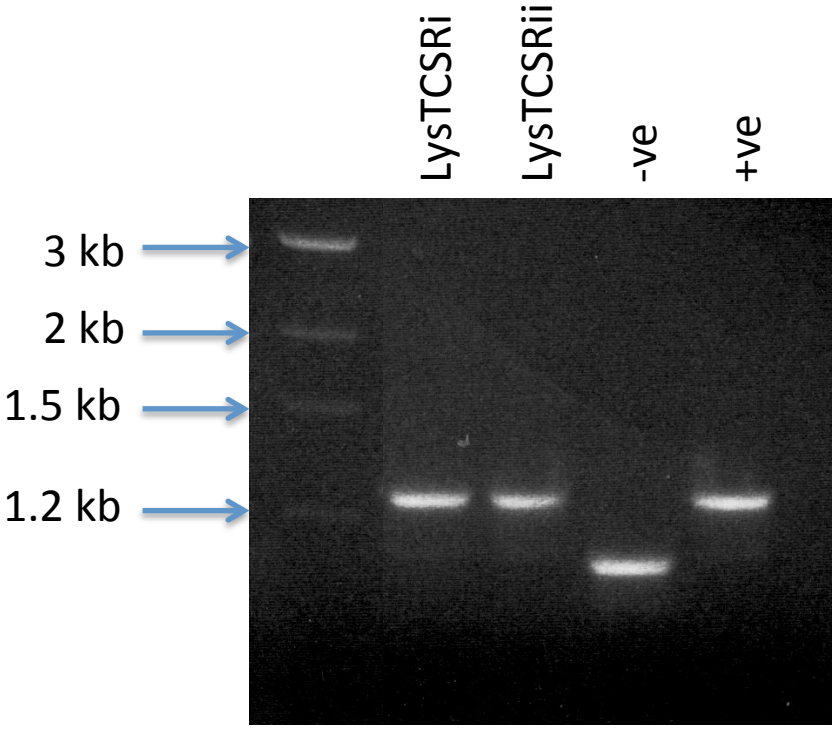
Appendix d– Confirmation of insertion of *cpl-1* into pSRSap1 by *Pvu*II digestion

Test digestion with *Pvu*II gave band patterns consistent with predictions for the creation of pSRSap1.cpl-1



Appendix e - Screening of putative TN72 transformants for correct insertion of *cpl-1* (*psaA* promoter)

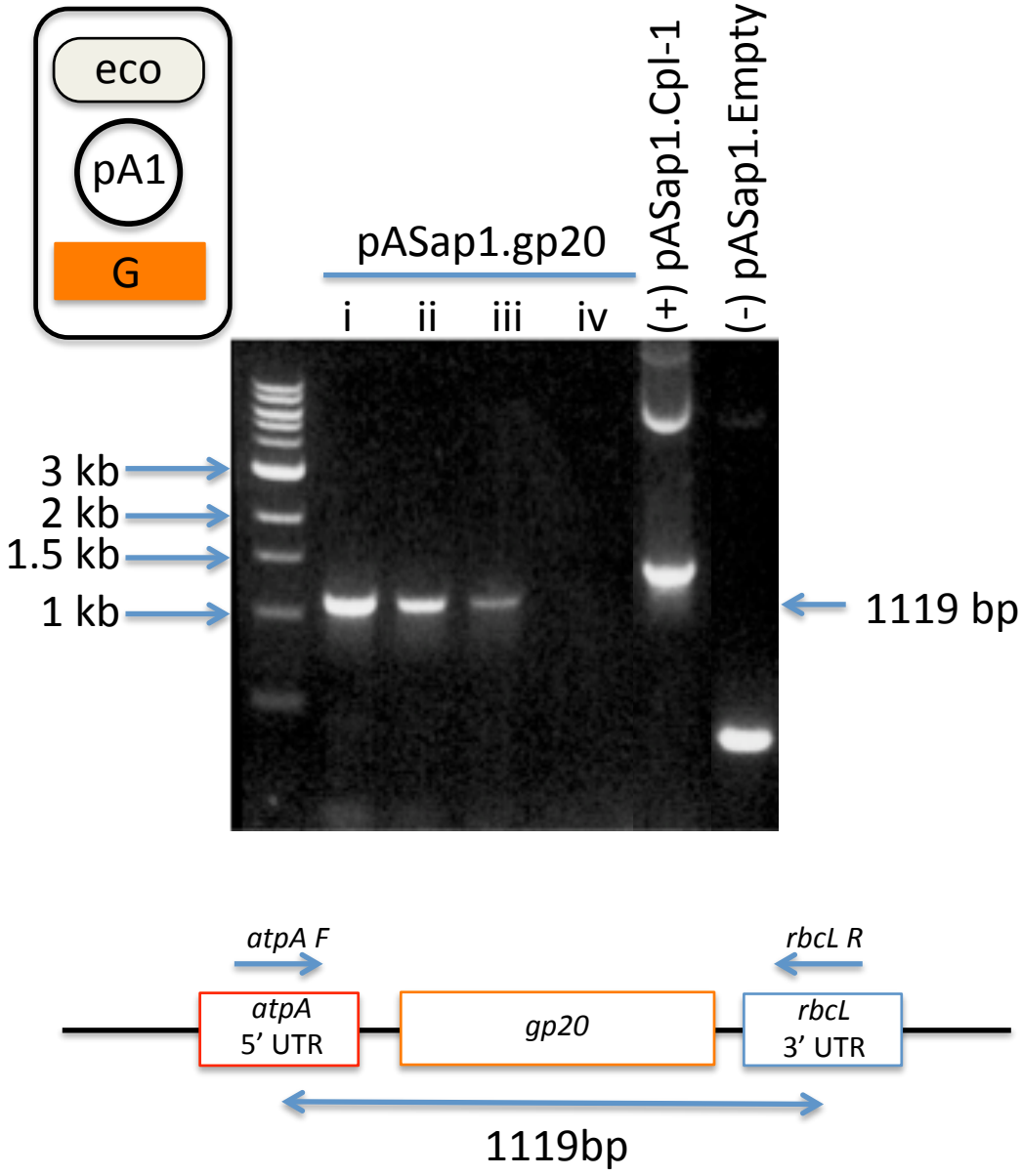
Two transformant colonies were re-streaked on HSM and screened for the correct insertion of the expression cassette. Both were confirmed as correct transformants.



Chapter four

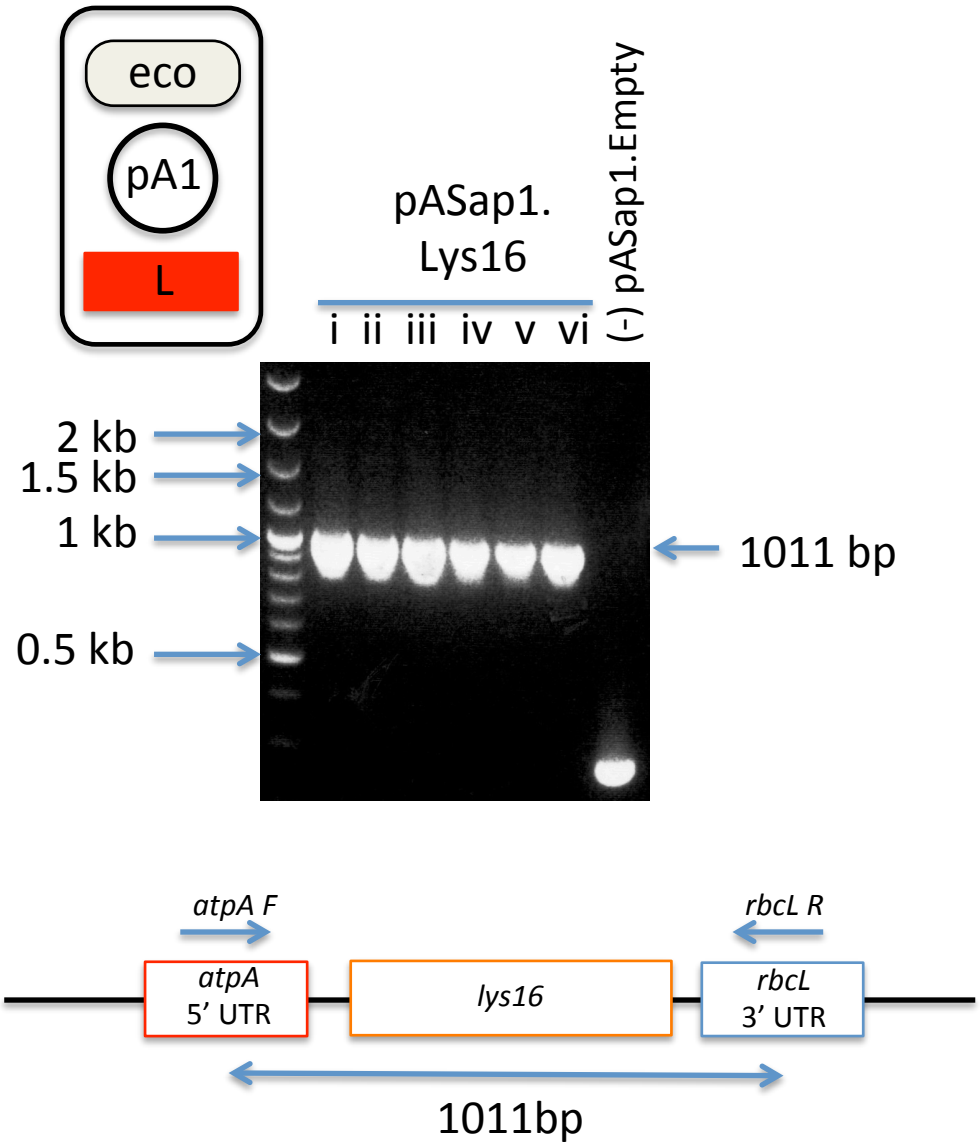
Appendix f – PCR confirmation of *gp20* insertion into pASap1

Insertion of the *gp20* gene into the pASap1 plasmid in *E. coli DH5α* was confirmed by colony PCR between the *atpA.F* and *rbcL.R* primers with the correct sized band running at 1119 bp.



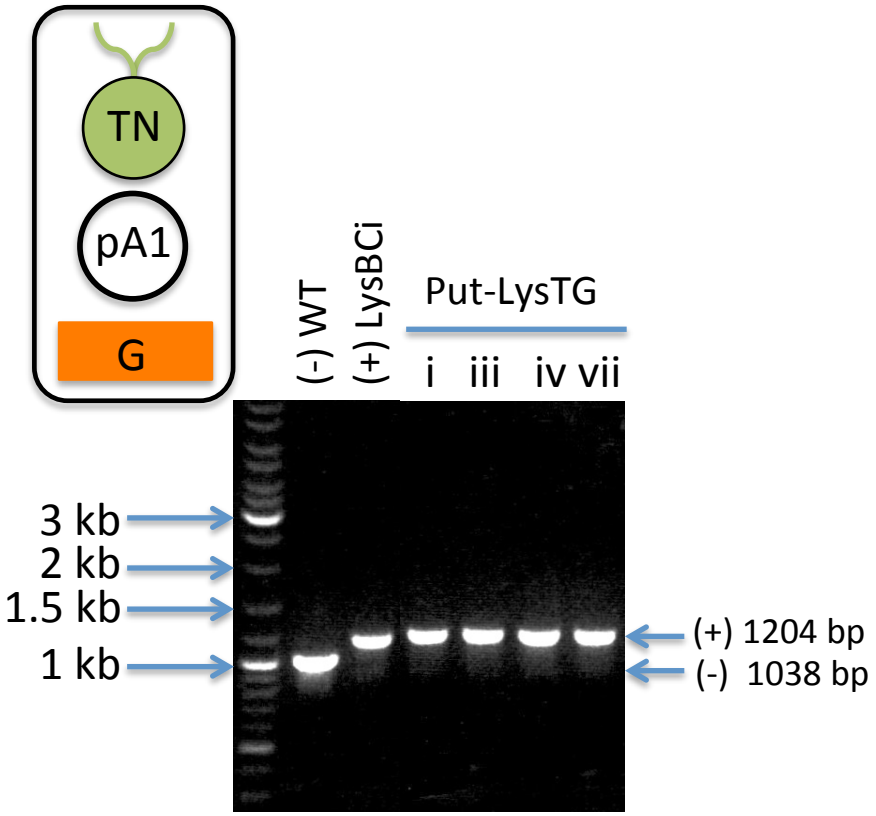
Appendix g - PCR confirmation of *lys16* insertion into pASap1

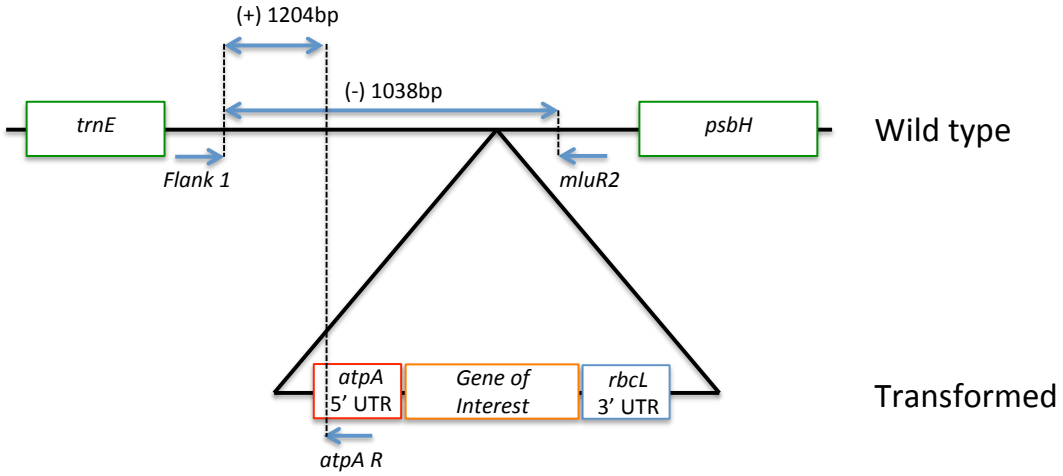
Insertion of the *lys16* gene into the pASap1 plasmid in *E. coli DH5α* was confirmed by colony PCR between the *atpA.F* and *rbcL.R* primers with the correct sized band running at 1011 bp.



Appendix h – PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pASap1.gp20

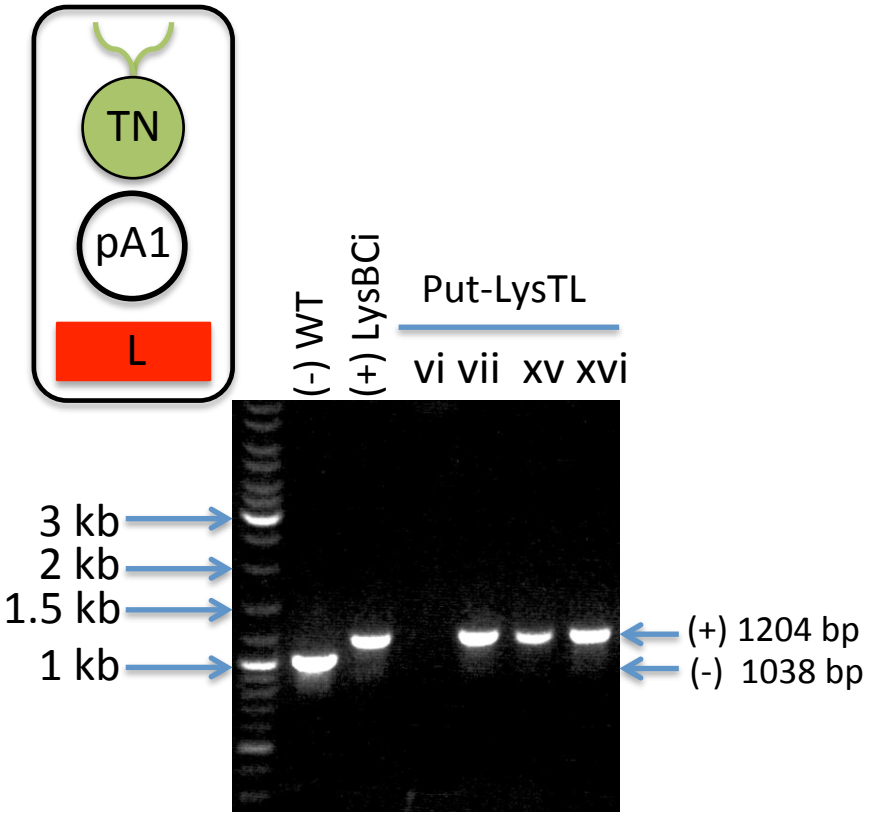
Putative transformant lines were screened by PCR following the schematic shown below. Correct insertion is indicated by a 1204 bp band. For this assay the putative transformants are screened relative to a wild type negative control.

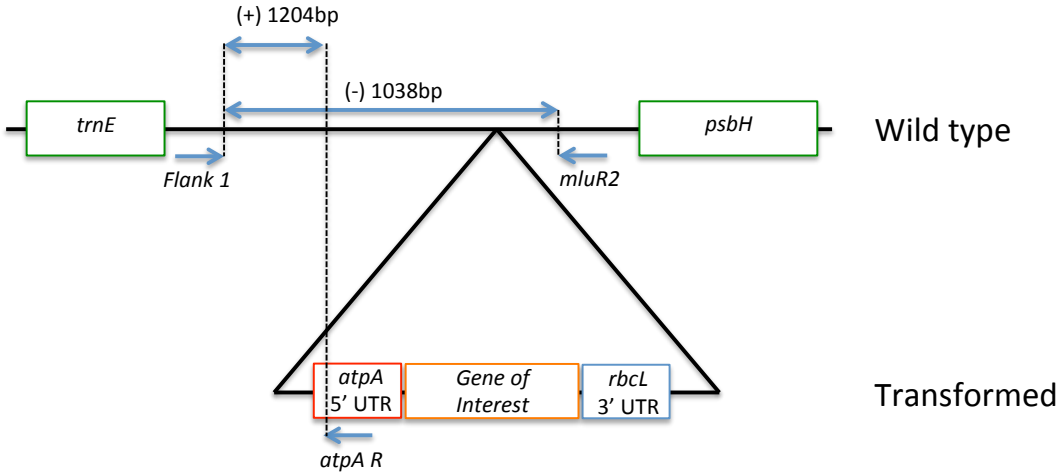




Appendix i – PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pASap1.lys16

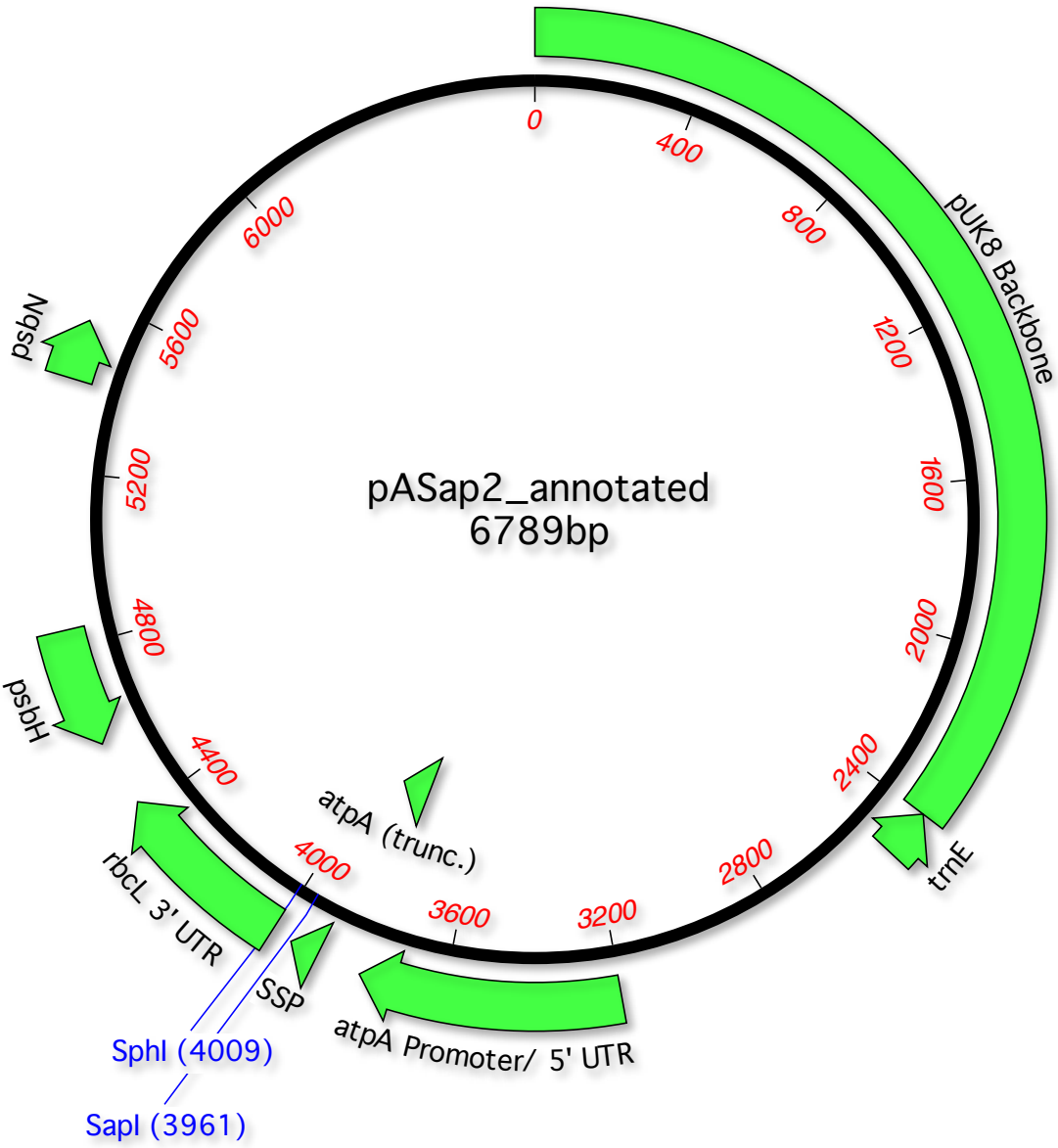
Putative transformant lines were screened by PCR following the schematic shown below. Correct insertion is indicated by a 1204 bp band. For this assay the putative transformants are screened relative to a wild type negative control.





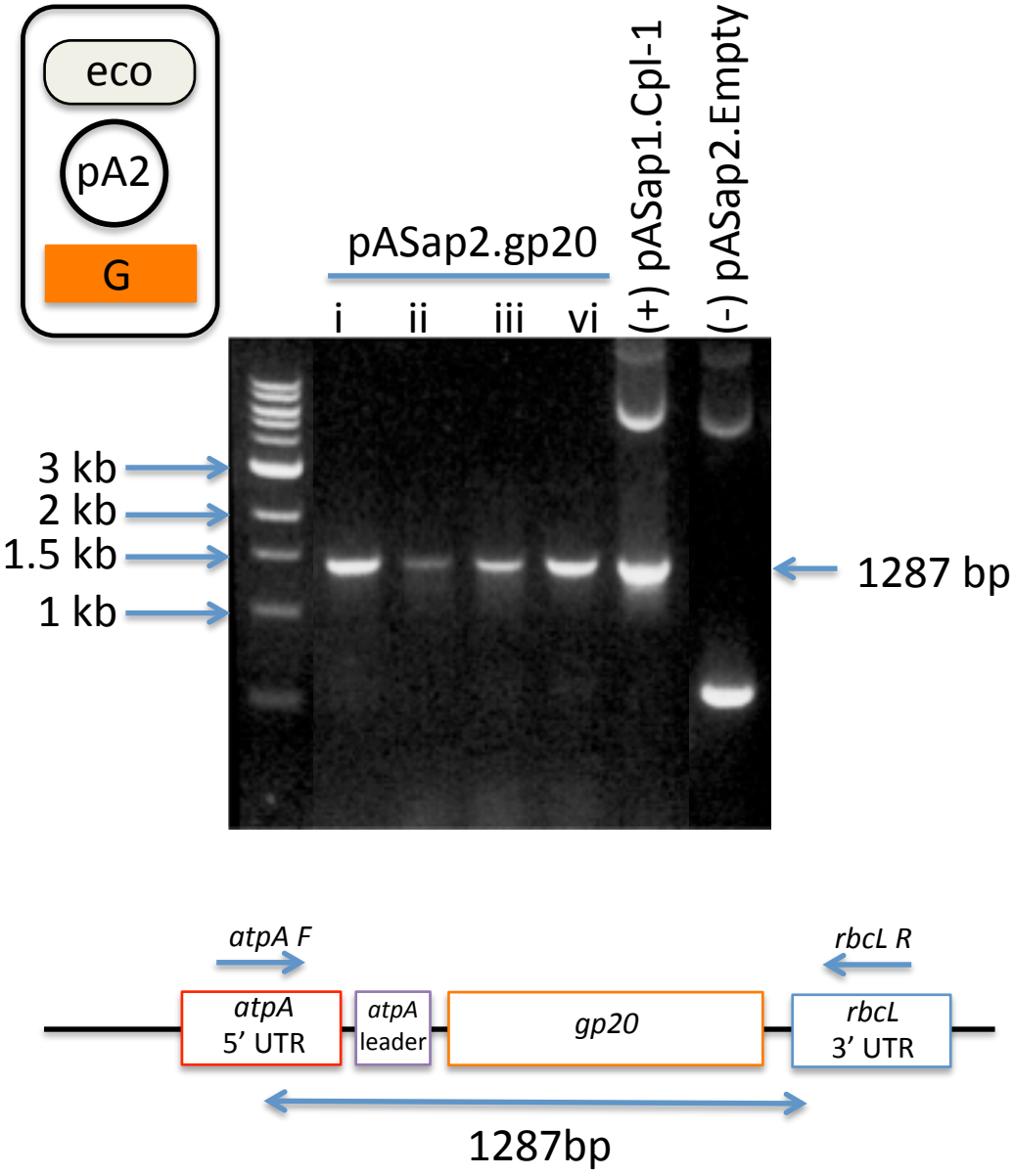
Appendix j – The pASap2 vector

The pASap2 transformation and expression vector is based on the pASap1 vector (see Figure 3.3), with the addition of the first 102 nucleotides of the endogenous *atpA* gene to the 5' end of the expression ORF. The stromal processing peptidase (SPP) site has also been inserted between the truncated *atpA* and the GoI so to allow for self-cleavage *in situ*.



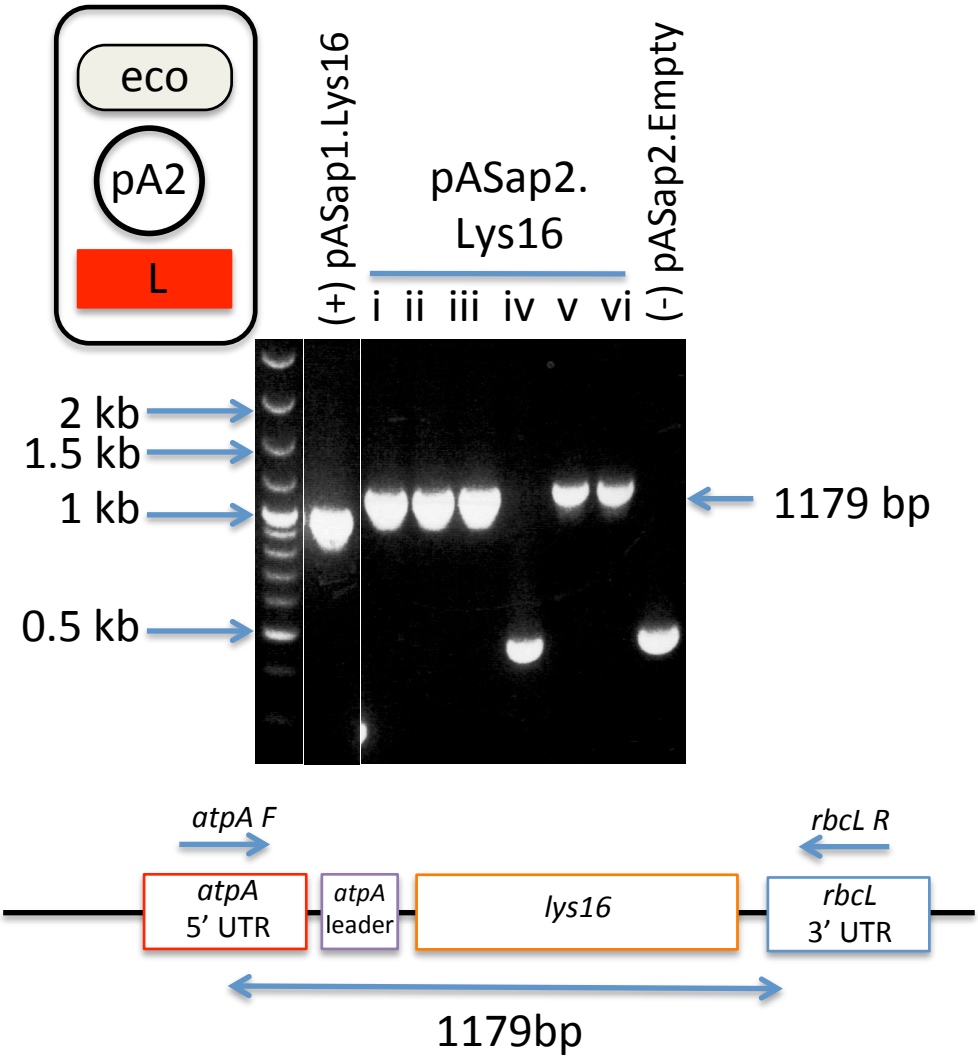
Appendix k – PCR confirmation of *gp20* insertion into pASap2

Insertion of the *gp20* gene into the pASap2 plasmid in *E. coli DH5α* was confirmed by colony PCR between the *atpA.F* and *rbcL.R* primers with the correct sized band running at 1287 bp.



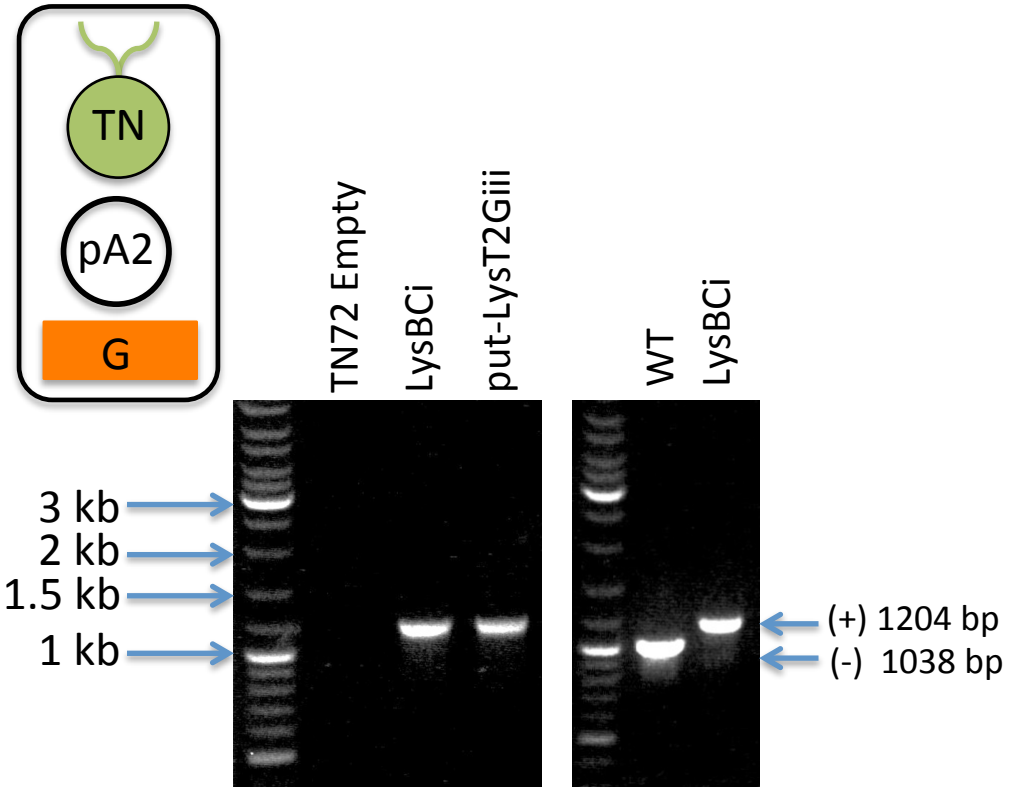
Appendix l – PCR confirmation of *lys16* insertion into pASap2

Insertion of the *lys16* gene into the pASap2 plasmid in *E. coli* DH5α was confirmed by colony PCR between the *atpA.F* and *rbcL.R* primers with the correct sized band running at 1179 bp.



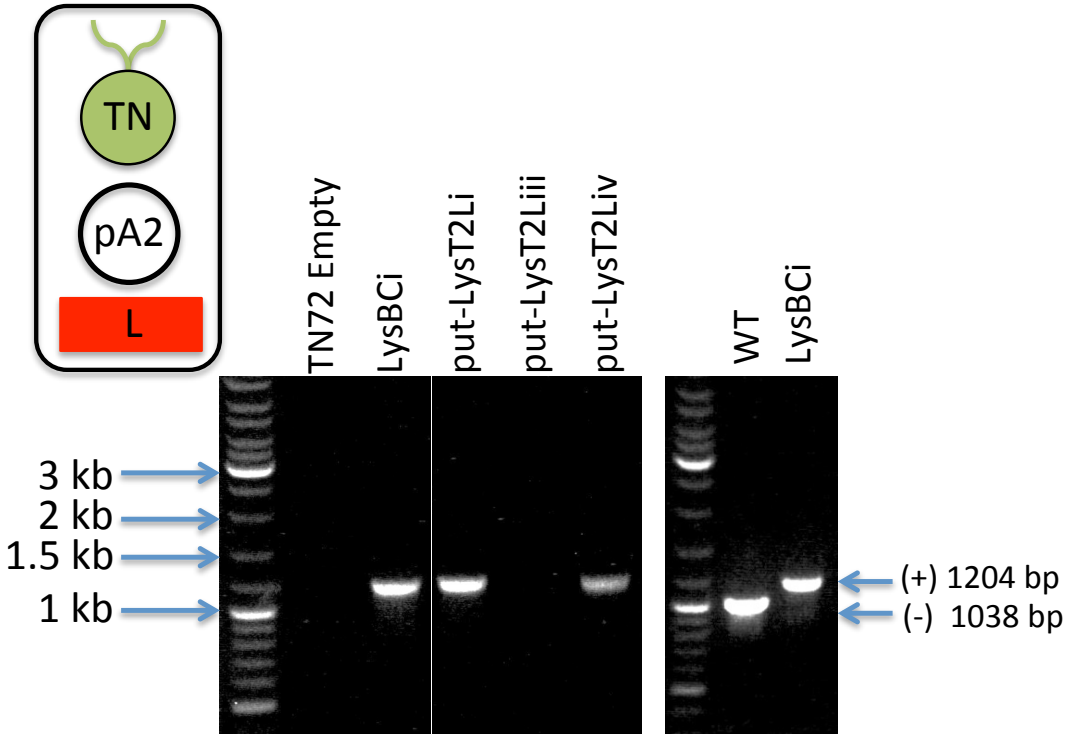
Appendix m – PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pASap2.gp20

The single putative transformant line was screened by PCR following the schematic shown in Appendix 4-c. Correct insertion is indicated by a 1204 bp band. As the TN72 negative control failed to amplify, a comparison of the positive control against wild type is provided.



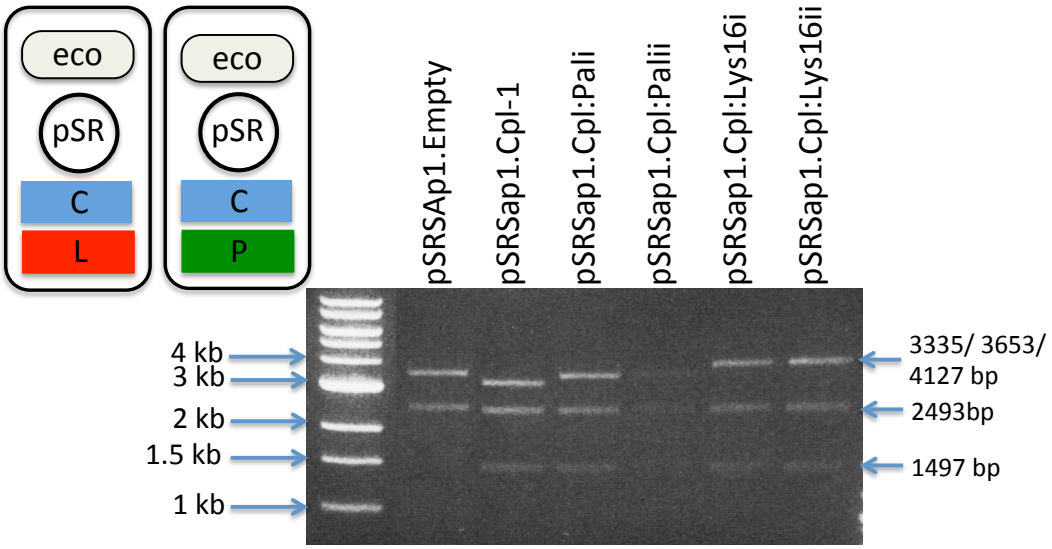
Appendix n – PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pASap2.lys16

Putative transformant lines were screened by PCR following the schematic shown above. Correct insertion is indicated by a 1204 bp band. As the TN72 negative control failed to amplify, a comparison of the positive control against wild type is provided.



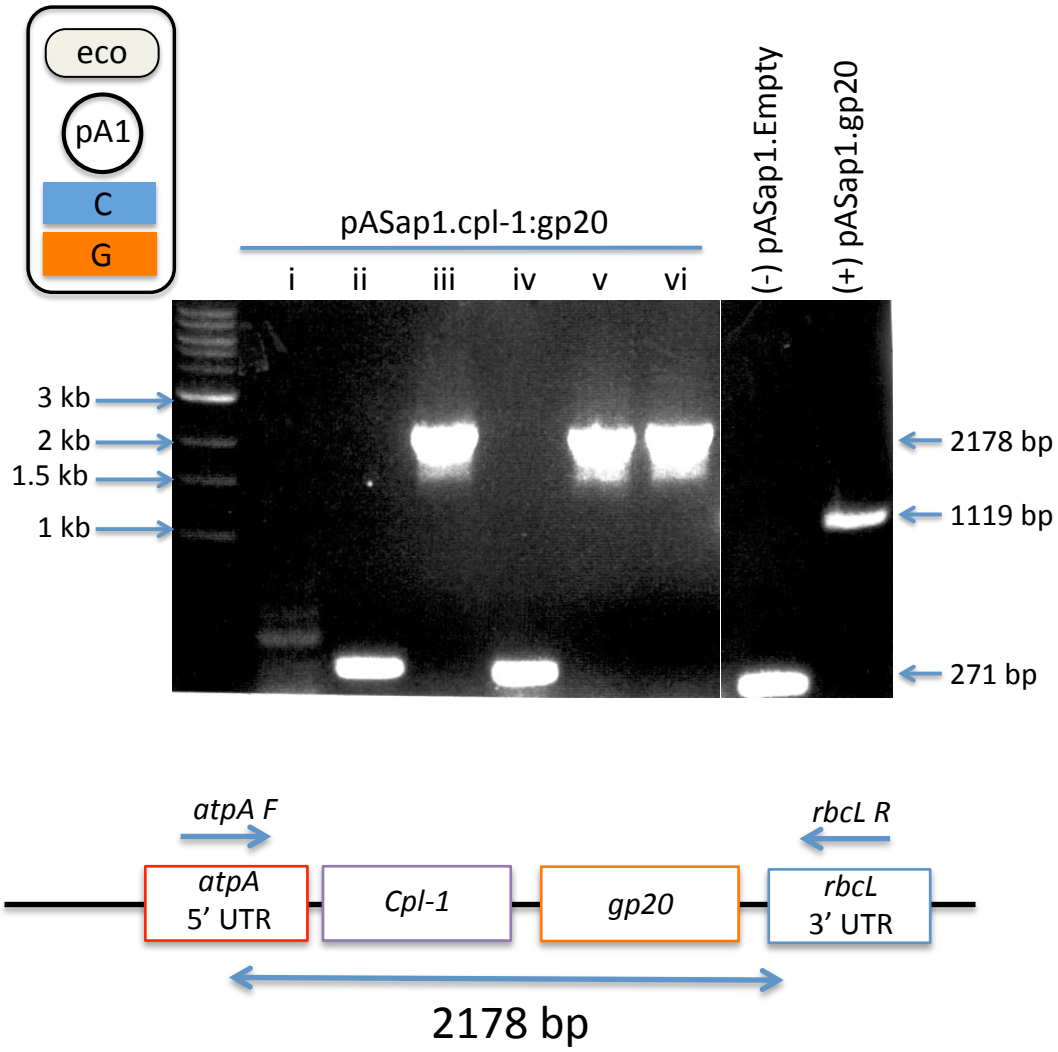
Appendix o – Insertion of *cpl-1, cpl-1:pal* and *cpl-1:lys16* into the pSRSap1 vector

Constructs were test-digested with *Pvu*II. All insertions featuring *cpl-1* give bands at 1497 and 2493 bp, with a third band at 3335, 3653, or 4127 bp indicating pSRSap1.cpl-1, pSRSap1.cpl-1:pal, or pSRSap1.cpl-1:lys16 respectively. pSRSap1.cpl-1:pal also yields bands at 597 and 15 bp (not shown).



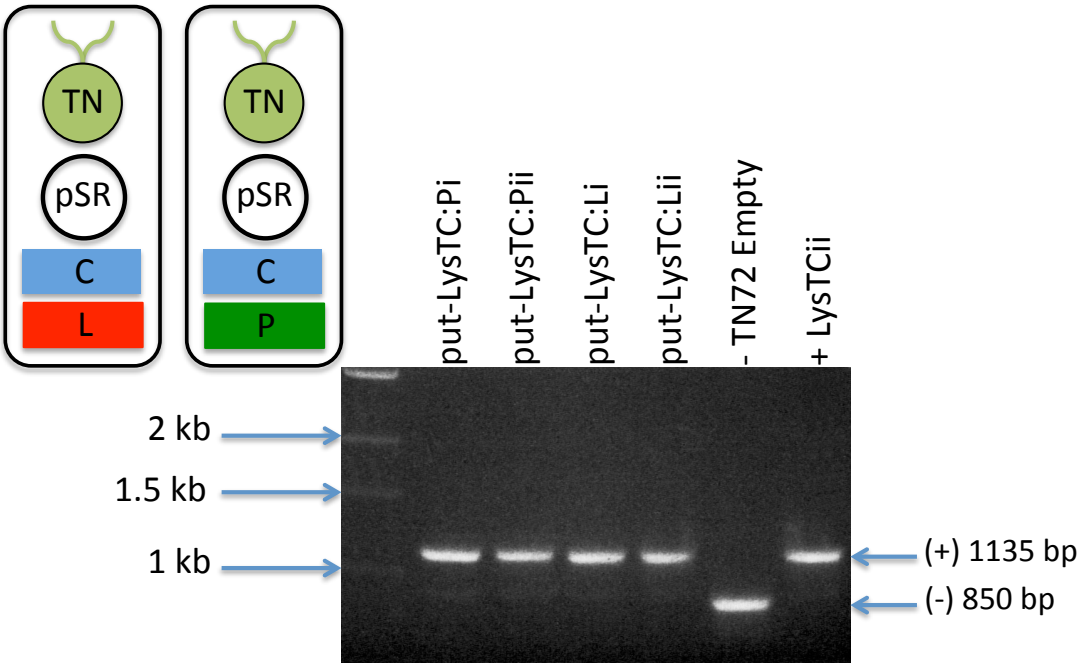
Appendix p - PCR confirmation of *cpl-1:gp20* insertion into pASap1

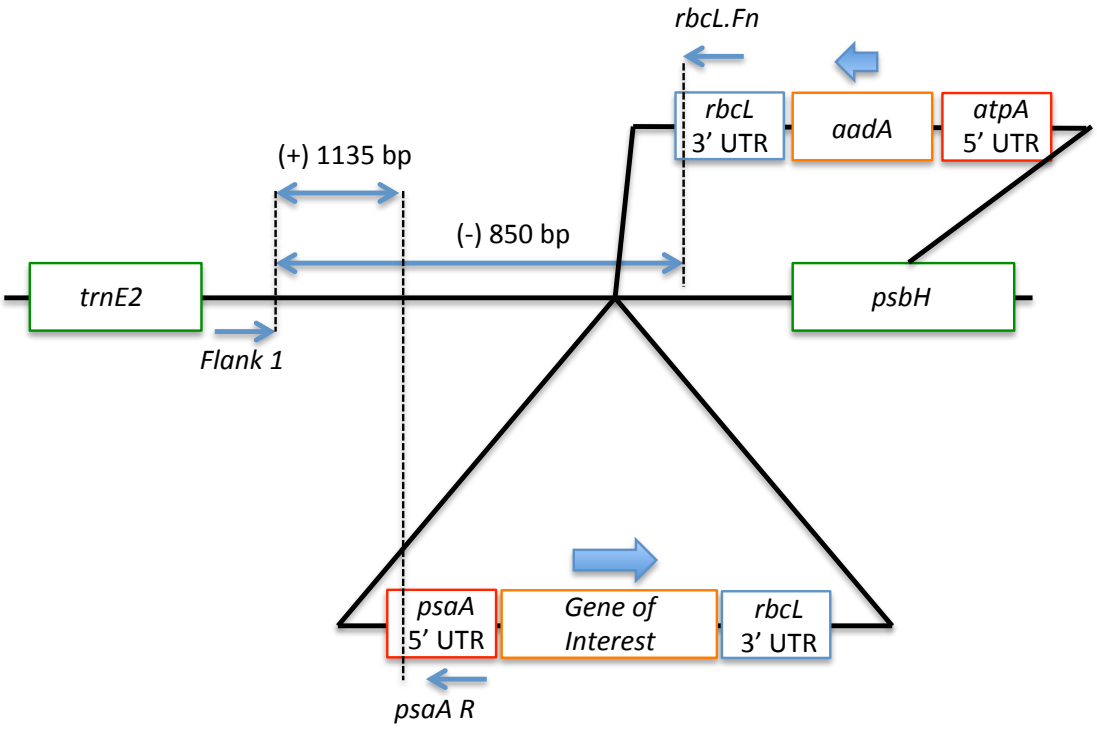
PCR between the *atpA* and *rbcL* primers gives products at 2178, 1119, and 271 bp for the full fusion, *gp20* only, and the empty vector respectively. This gel shows correct insertion for 3 of the six constructs, with 2 showing the empty vector, and 1 giving a weak band at approximately 400 bp, possibly from a random insertion.



Appendix q – PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pSRSap1.cpl-1:lys16 and pSRSap1.cpl-1:pal

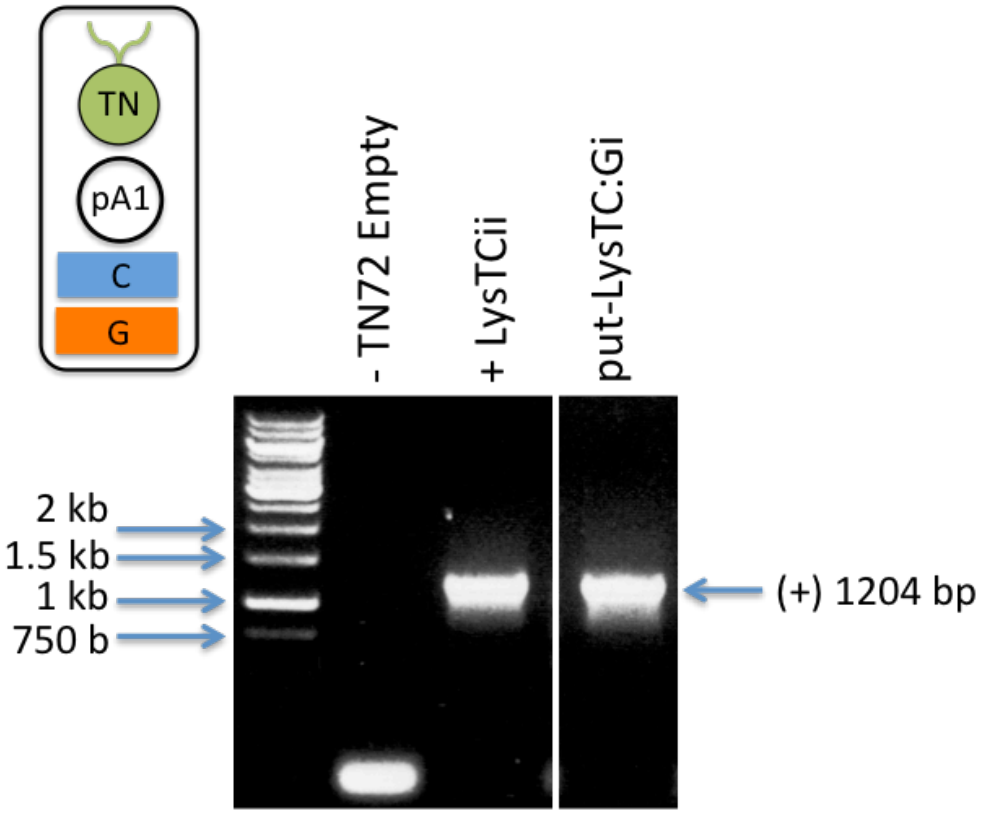
PCR shows correct transformation of all four putative lines. Low levels of heteroplasmy are observed.





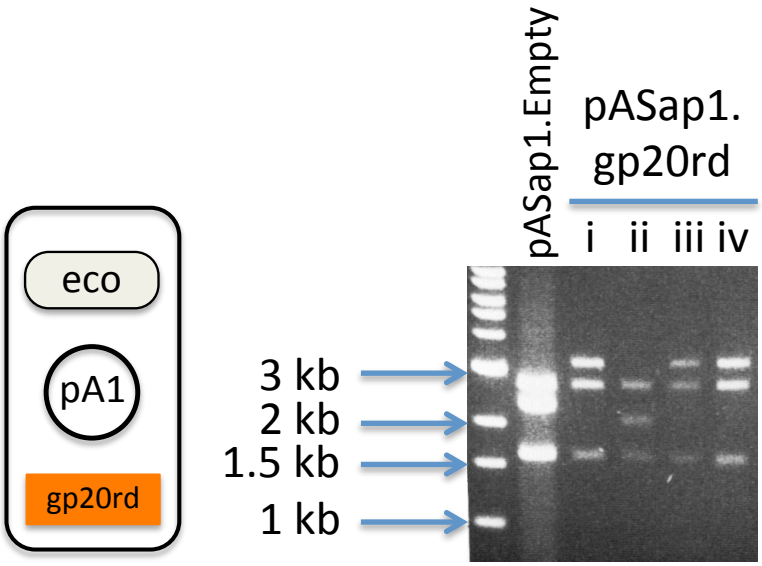
Appendix r - PCR confirmation of transformation of the *C. reinhardtii* recipient line TN72 with pASap1.cpl-1:gp20

The putative line LysTC:Gi is shown to contain the pASap1.cpl-1:gp20 insert giving a band at 1204 bp.



Appendix s - Insertion of *gp20rd* into the pASap1 vector

Confirmation of insertion is by test digest with *XmnI.* All samples give bands at 1601 and 2672 bp, with correct insertions containing a third band at 3100 bp as opposed to 2348 bp for the empty vector. Samples i, iii, and iv are shown to contain the insert.



Appendix t – Primers used during Chapters three and four

Name	Purpose	Sequence
PA6.Int'nal.F	Internal sequencing primers for *gp20*	agtttggccagctgttgaacgtg
PA6.Int'nal.R		cacaaacaccacgacgacctgc
L16.Int'nal.F	Internal sequencing primers for *lys16*	atgtttagaacaaaactggttaggtg
L16.Int'nal.F		ccgtattggttacgtttccatttac
atpA.F	pASap1/2 sequencing primers	caagtgatcttaccactcac
rbcL.R		gctgctgcatgtgaagtttg
psa.F	pSRSap1 sequencing primers	gttcacgcgtaagctttcttaattcaacattt
rbcL.R		gctgctgcatgtgaagtttg
Flank1	Screening primers for pASap1/ 2 transformants	gtcattgcgaaaatactggtgc
atpA.R		acgtccacaggcgtcgtaagc
mluR2		gatgacgtttctatgagttggg
rbcL.Fn	Screening primers for pSRSap1 transformants	cggatgtaactcaatcggtag
psaA.R		ggatttctccttataataac
cpl-1Fus.F	Fusion lysin construction	gctagctcttctatggttaaaaaaaatgatttattcgttg
cpl-1Fus.R		catcgaggaaccaccaccacctccacctgatgagccaccaccaccagcaactgtaattaaaccatctg
pal Fus.F		atgggtgttgatattgaaaaagg
L16.Fus.F		ggtggtggtggctcatcaggtggaggtggtggtggttcctcgatgaaatcacaacaacaagctaaa
gp20.Fus.F		ggtggtggtggctcatcaggtggaggtggtggtggttcctcgatggttcgttatatatccagct
pal.Fus.R		gcttgcatgcttattaagcg
L16.Fus.R		gcttgcatgcttattaagcata

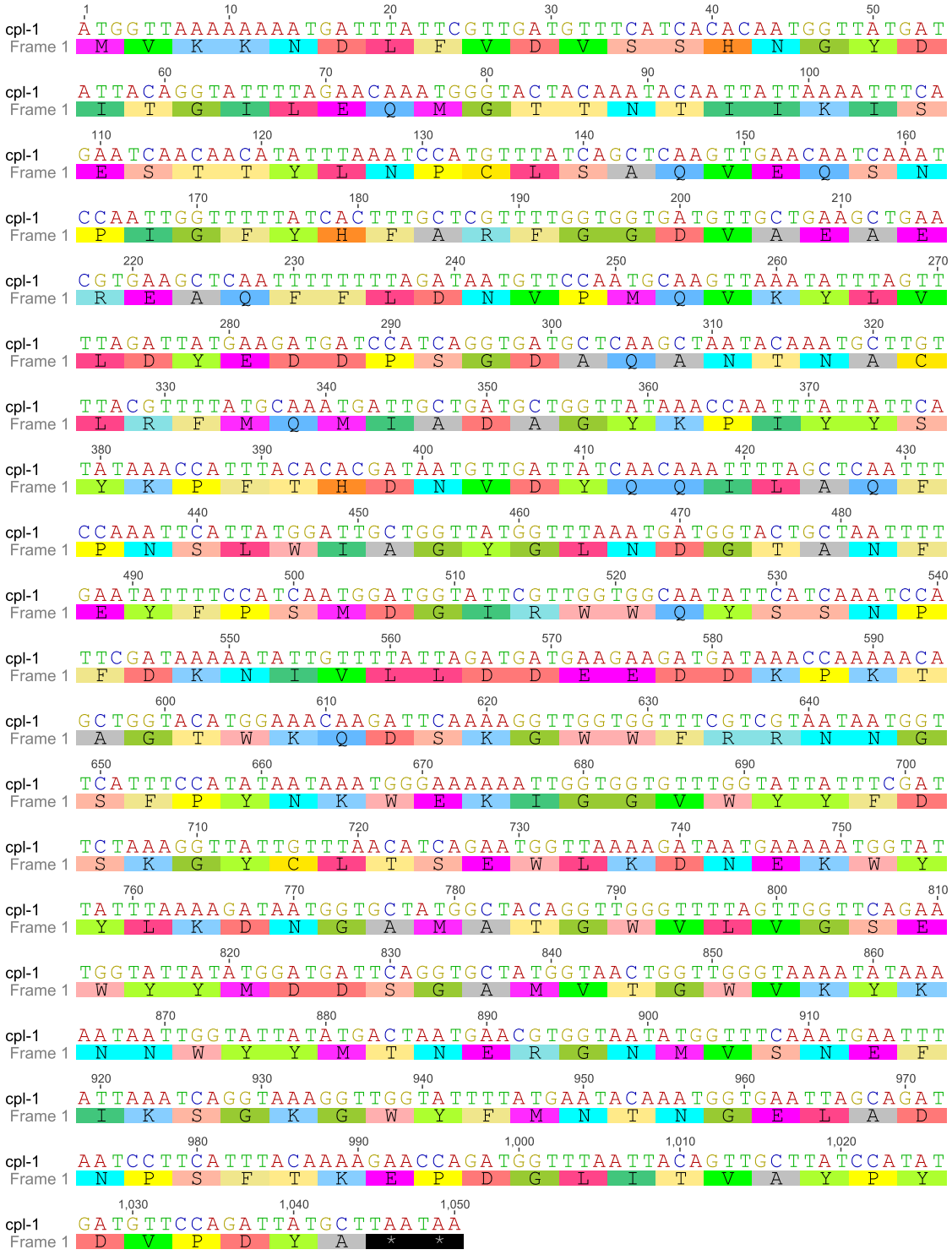
`Chapter five

Appendix u – The observed and predicted datasets directly analysed in Chapter five

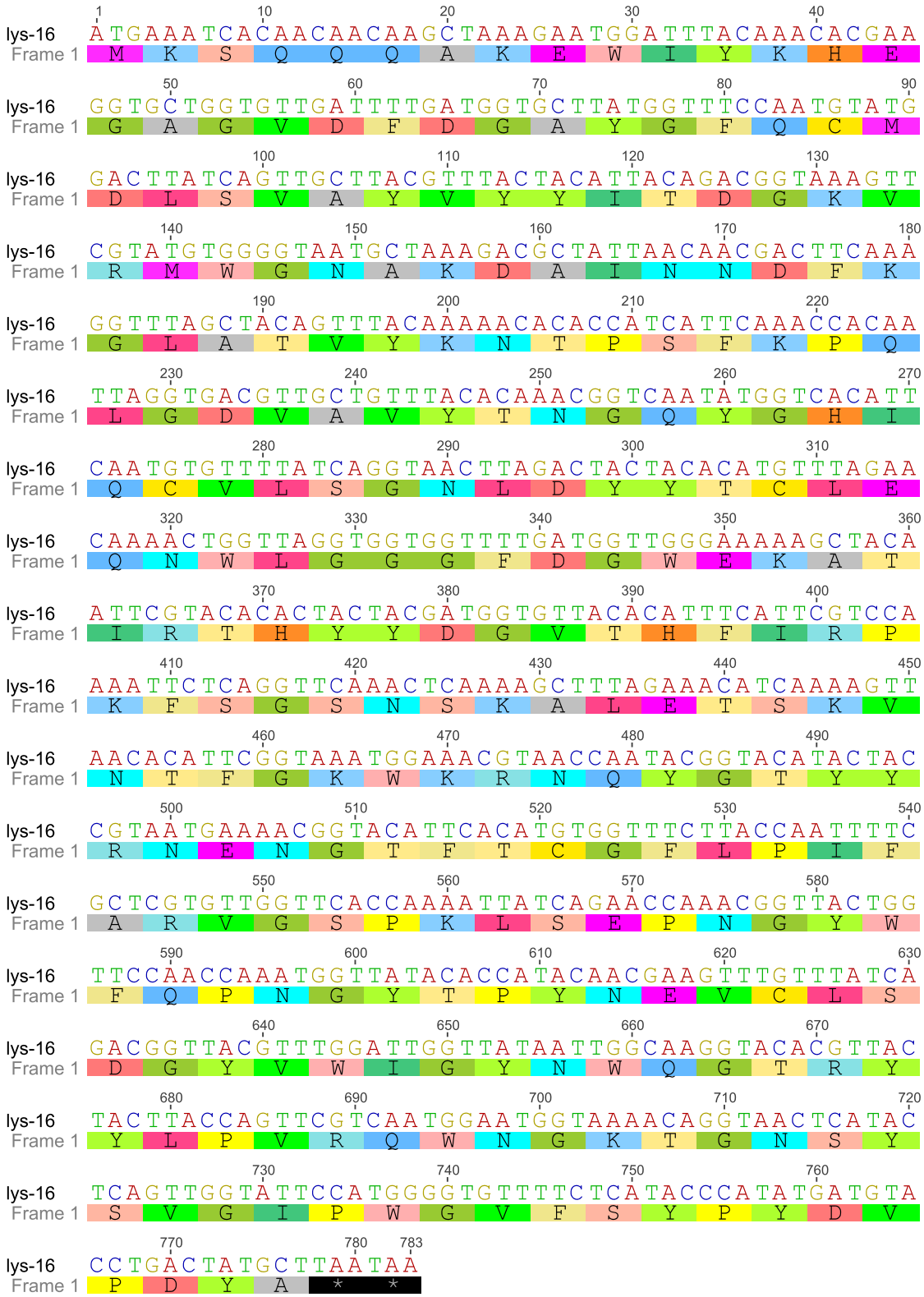
Dataset	Features
o(datasset1)	The observed codon- and codon pair usage for the 68 unique protein-coding genes in the *C. reinhardtii* chloroplast genome
p(dataset1)	Codon distribution predictions based on amino acid usage alone
p(dataset2)	Codon pair distribution predictions based on observed absolute frequencies of individual codons
$p(dataset2\alpha)$	Back translation of *p(dataset2)* to allow an amino acid pair comparison to *o(dataset1)*
p(dataset3)	Codon pair distribution predictions based on observed relative frequencies of individual codons and amino acid pair usage
e(dataset1)	Codon- and codon pair distributions observed in the 11 expressing transgenes listed in Table 5.2
n(dataset1)	Codon- and codon pair distributions observed in the 11 non-expressing transgenes listed in Table 5.2

Sequences

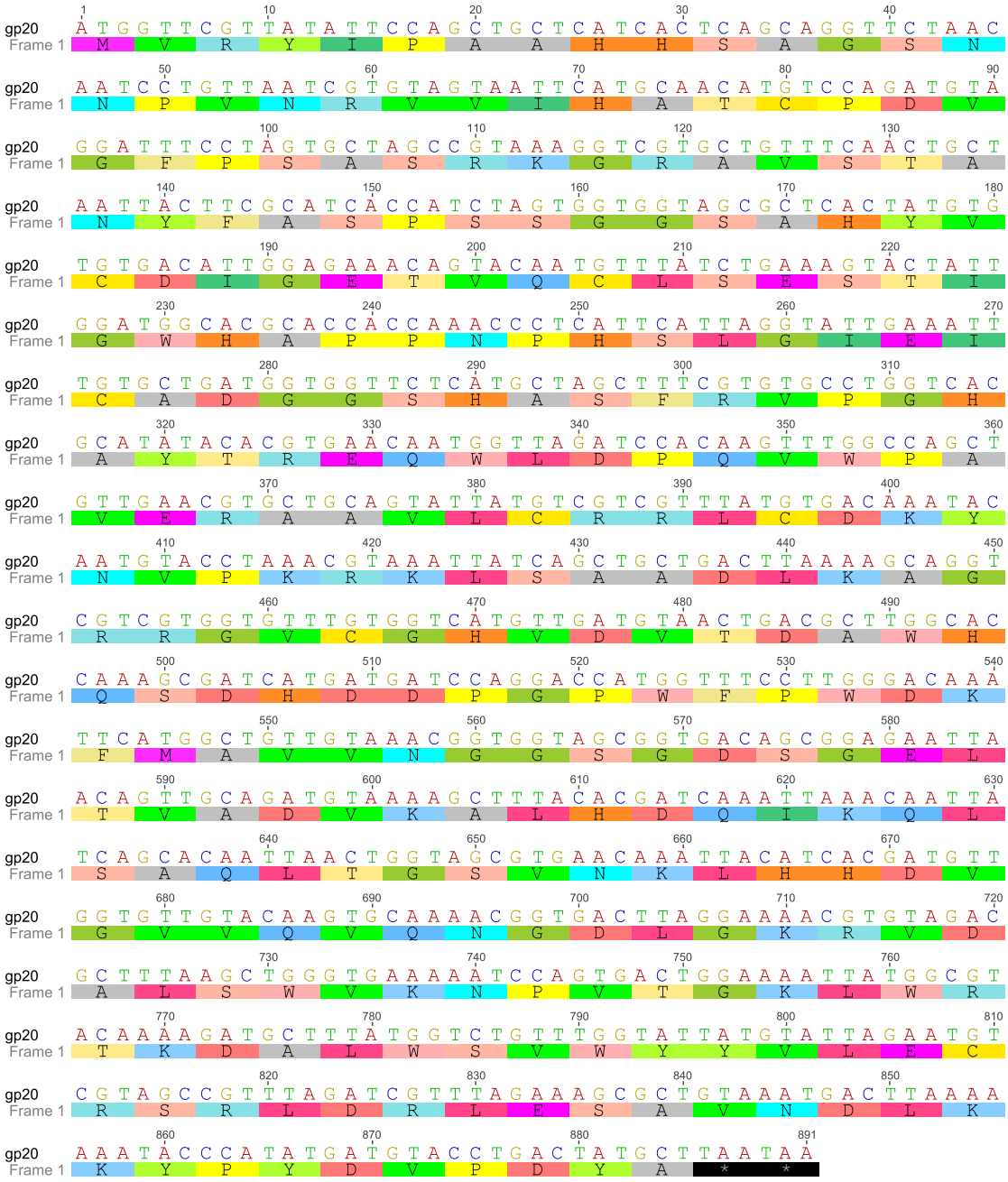
Appendix v – Gene sequence of *cpl-1* with translation



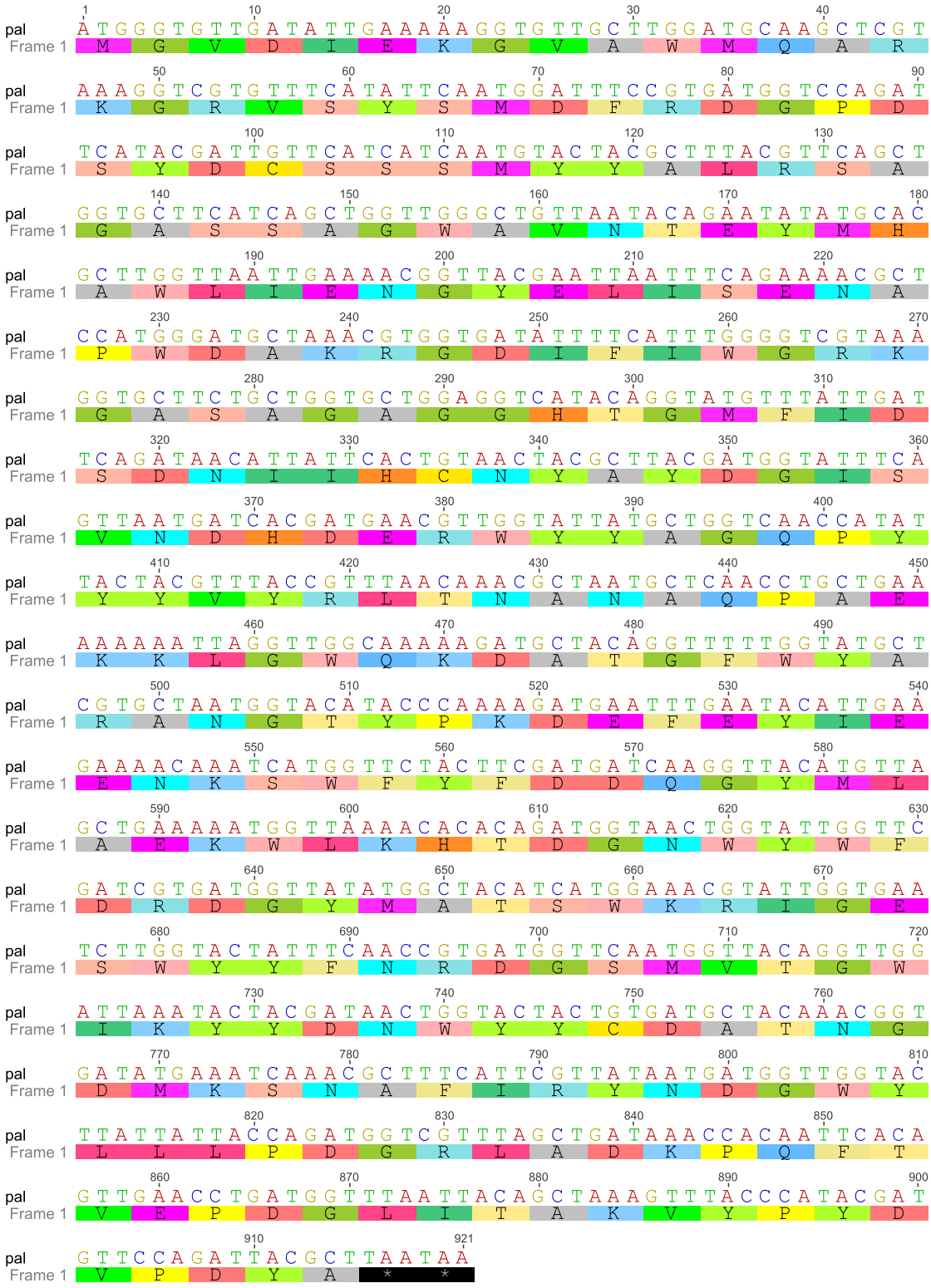
Appendix w - Gene sequence of *lys-16* with translation



Appendix x - Gene sequence of *gp20* with translation

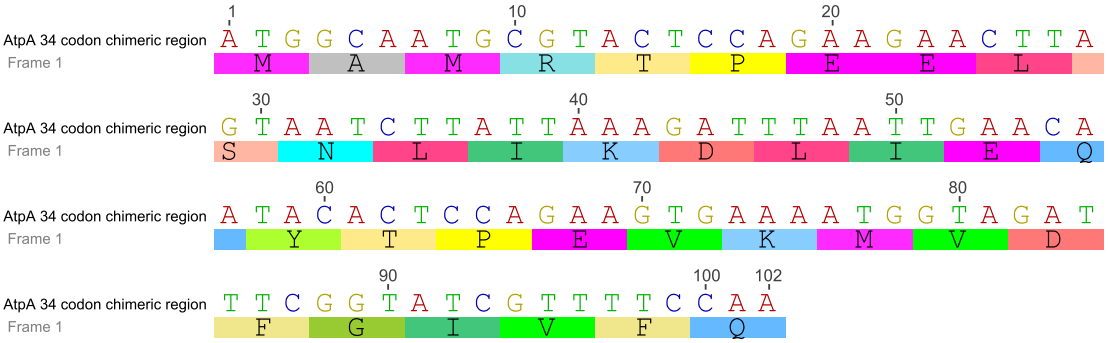


Appendix y – Gene sequence of *pal* with translation

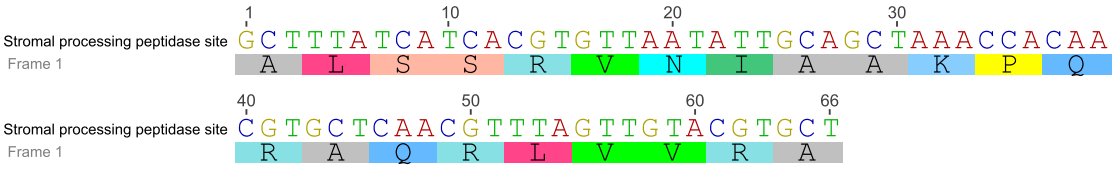


Appendix z – Gene sequences of chimeric and fusion protein linker regions with translations

AtpA 34 codon chimeric region



Stromal processing peptidase site



Fusion construct flexi-linker